

## LOG REKAM JEJAK KORESPONDENSI PUBLIKASI ILMIAH

### Syarat Khusus Usulan Jabatan Fungsional Lektor Kepala

#### A. Identitas Karya Ilmiah

Judul Artikel : Development and Comparative Evaluation of Machine Learning Models using Clinically Relevant Features for Predicting Newborn Patients' Length of Stay

Nama Jurnal : Jurnal Teknik Informatika (JUTIF)

Volume/No/Tahun : Vol. 7/No. 3/June 2026

Peringkat Akreditasi : Sinta 2

Status Penulis : Penulis Pertama sekaligus Penulis Koresponden

Tautan (URL) Artikel : <https://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/5410/1261>

#### B. Kronologi Aktivitas Korespondensi (Peer Review Log)

| No. | Tanggal    | Tahapan Aktivitas               | Keterangan / Catatan Tim Editor & Reviewer                                                                                        | Lampiran                                                                                                                                                                                                                                                                          | Halaman |
|-----|------------|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| 1.  | 10/10/2025 | <i>Submission of Manuscript</i> | Artikel pertama kali diunggah ke sistem OJS jurnal. Pengusul menerima email konfirmasi <i>Submission Received</i> .               | Lampiran 1: Bukti pengiriman di OJS JUTIF<br>Lampiran 2: Bukti pengiriman di Email                                                                                                                                                                                                | 3-4     |
| 2.  | 4/12/2025  | <i>Revision Required</i>        | Menerima Keputusan dari Editor dan Menerima komentar dari Reviewer 1 dan Reviewer 2.<br><br>Status: Jurnal diterima dengan Revisi | Lampiran 3: Screenshot halaman review di OJS JUTIF<br>Lampiran 4: Keputusan Hasil Review dan Editor dikirim via email Email<br>Lampiran 5: Keputusan Hasil Review dan Editor di OJS JUTIF<br>Lampiran 6: Hasil Review dari Reviewer 1<br>Lampiran 7: Hasil Review dari Reviewer 2 | 5-52    |

|    |            |                                  |                                                                                                                                                                                                                                                         |                                                |         |
|----|------------|----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|---------|
| 3. | 15/12/2025 | <i>Submit Revised Manuscript</i> | Pengusul mengunggah artikel perbaikan beserta dokumen <i>Response to Reviewers</i> (jawaban revisi).<br><br>Hasil Perbaikan pada artikel diberikan tanda warna kuning.<br><br>Artikel diperbaiki sesuai dengan masukan/review Reviewer 1 dan Reviewer 2 | Lampiran 8: bukti artikel yang sudah direvisi  | 53-71   |
| 4. | 5/1/2026   | <i>Invoice Notification</i>      | Bukti artikel diterima setelah revisi dari Editor-in-Chief dan tagihan pembayaran APC, proses selanjutnya adalah pembayaran                                                                                                                             | Lampiran 9: Bukti Penerimaan Arsikel via email | 72-73   |
| 5. | 19/1/2026  | <i>Acceptance Notification</i>   | Surat Penerimaan Artikel ( <i>Letter of Acceptance / LoA</i> ) dari Editor-in-Chief.                                                                                                                                                                    | Lampiran 10: Bukti LoA                         | 74      |
| 6. | 15/6/2026  | <i>Article Published Online</i>  | Artikel resmi terbit pada Volume 7, Nomor 3, Tahun June 2026                                                                                                                                                                                            | Lampiran 11: Bukti Penerbitan Artikel          | 75 - 82 |

Demikian rekam jejak korespondensi ini disusun dengan sebenarnya untuk melengkapi berkas syarat khusus usulan Lektor Kepala bidang Ilmu Komputer.

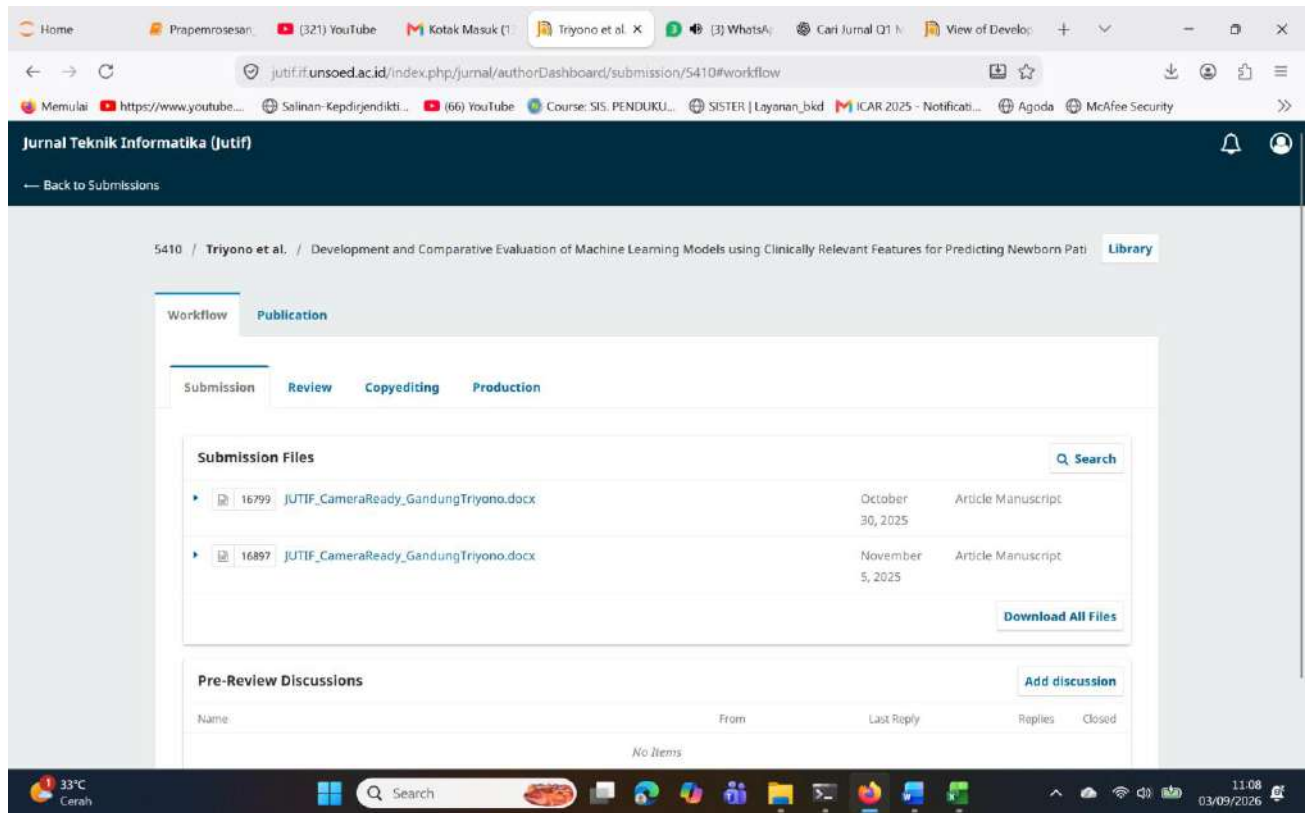
Jakarta, 1 September 2026  
Pengusul,



Dr. Ir. Gandung Triyono, S.Kom., M.Kom.  
0306027701/970020

# LAMPIRAN-LAMPIRAN

## 1. Bukti Submission Acknowledgemen di OJS



## 2. Bukti Submission Acknowledgemen di Email



Gandung Triyono <gandung.triyono@budiluhur.ac.id>

---

### [Jutif] Submission Acknowledgement

1 pesan

---

**JUTIF UNSOED** <jutif.ft@unsoed.ac.id>

30 Oktober 2025 pukul 12.13

Kepada: Gandung Triyono <gandung.triyono@budiluhur.ac.id>

Gandung Triyono:

Thank you for submitting the manuscript, "Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay" to Jurnal Teknik Informatika (Jutif), with e-ISSN: 2723-3871 and is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024.

With the online journal management system that we are using, you will be able to track its progress through the editorial process by logging in to the journal web site:

Submission URL: <https://jutif.if.unsoed.ac.id/index.php/jurnal/authorDashboard/submission/5410>

Username: gandungtriyono

If you want take fasttrack, please contact : 085640661444 for detail.

Thank you for considering this journal as a venue for your work.

JUTIF UNSOED

---

Jurnal Teknik Informatika (Jutif)

3. Keputusan Editor di OJS

Jurnal Teknik Informatika (Jutif)



← Back to Submissions

5410 / Triyono et al. / Development and Comparative Evaluation of Machine Learning Models u

Library

Workflow

Publication

Submission

Review

Copyediting

Production

Round 1

Round 1 Status

Submission accepted.

Notifications

|                                         |                     |
|-----------------------------------------|---------------------|
| <a href="#">[jutif] Editor Decision</a> | 2025-12-04 06:37 AM |
| <a href="#">[jutif] Editor Decision</a> | 2026-01-05 04:57 AM |
| <a href="#">[jutif] Editor Decision</a> | 2026-01-05 12:42 PM |

Reviewer's Attachments

Q Search

|                                                                                     |       |                                                                                       |                   |
|-------------------------------------------------------------------------------------|-------|---------------------------------------------------------------------------------------|-------------------|
|  | 17278 | <a href="#">jutif-review-assignment-5410-Article+Manuscript-16898 - Reviewed.docx</a> | November 26, 2025 |
|  | 17510 | <a href="#">jutif-review-assignment-5410-Article+Manuscript-16898_rev.docx</a>        | December 4, 2025  |

Revisions

Q Search

Upload File

|   |                                                                                     |       |                                     |                   |                    |
|---|-------------------------------------------------------------------------------------|-------|-------------------------------------|-------------------|--------------------|
| ▶ |  | 17834 | <a href="#">JUTIF-Template.docx</a> | December 15, 2025 | Article Manuscript |
|---|-------------------------------------------------------------------------------------|-------|-------------------------------------|-------------------|--------------------|

Review Discussions

Add discussion

| Name     | From | Last Reply | Replies | Closed |
|----------|------|------------|---------|--------|
| No Items |      |            |         |        |

## 4. Keputusan Hasil Review dan Editor di Email



Gandung Triyono <gandung.triyono@budiluhur.ac.id>

---

### [Jutif] Editor Decision

1 pesan

Nur Chasanah <jutif.ft@unsoed.ac.id>

4 Desember 2025 pukul 13.37

Kepada: Gandung Triyono <gandung.triyono@budiluhur.ac.id>

Gandung Triyono:

We have reached a decision regarding your submission to Jurnal Teknik Informatika (Jutif), with e-ISSN: 2723-3871 and is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024, with title "Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay".

Our decision is: Revisions Required

In addition, several things to consider from the editor's side are as follows :

1. Please **FOLLOW and USE** JUTIF's template. You can download it at <http://jutif.if.unsoed.ac.id/index.php/jurnal/about/submissions>. After that, you can copy your paper into JUTIF's template.
2. Please use table and figure format according to JUTIF's Template. The resolution of figures must be clear, can be seen and **more than 300 dpi**.
3. Every figure and table **MUST BE** cited and must be explained in paragraph, ex : "Research method can be seen in Figure 1", "Table 1 showed XYZ."
4. Section "Discussion" is **MANDATORY**. If your article did not contain Discussion, please add it.
5. **MINIMUM** references in JUTIF is **25 primary references** (Journal/Conference) from 5 years latest. Please add reference if necessary. In JUTIF, Reference using IEEE format. Please adjust in your reference. We recommend to use reference manager tools, such as : Mendeley, Zootero, or EndNote.
6. Please Correct the manuscript **according to the review** from the reviewer or editor. Comments from reviewers that need to be corrected for the manuscript has been moved at the bottom.
7. Please **give mark the revised results on the revised manuscript sent with YELLOW HIGHLIGHTS** to the corrected sentence or sentences added of revision fulfillment from the Reviewer.
8. Please upload revision file along with Turnitin Result File (If Any). The **Maximum plagiarism result is 20%**.
9. Revisions can be sent via the "File upload" in the "revision" section of the JUTIF's OJS. Please when upload the Revised File, do not overwrite the previous revised file (if any).
10. For Information, you can contact our representative : 0823-2492-4093.

Thank You.

Best Regards.

-----  
Reviewer A:  
Recommendation: Revisions Required  
-----

#### Title

Novelty not reflected in the title.  
Could add dataset context, domain, etc.

#### Abstract

Could add explicit gap statements and novel contributions.

#### Introduction

Could add a paragraph to explain the research gap.  
The literature review is descriptive; it should be more comparative.

#### Method

- Could add a figure for the research framework, showing:  
Data Collection → Preprocessing → Feature Engineering → Modeling → Evaluation
- Could add a table of dataset attributes.
- Fix equations formatting. Read: Jutif Template.
- In Methods, should focus on implementation, not just on mathematical algorithms.

#### Results

- Add more descriptive results.
  - The table format should follow Jutif style. Read: Jutif template.
- Download Reviewer Files for more details.

#### Discussions

- Add comparison with the previous study.
- Add more explanation of why RFR works better for this dataset.
- Add practical implications for Hospitals.
- Could discuss the limitations.

#### Conclusion

Add an explicit contribution statement.

## References

- Some DOIs are missing and the journal names are not in full format. Fix it.
- The WHO website citation should not dominate.
- Correct spelling errors in the references.
- Preprints references should be minimized.

## Overall Comment

This manuscript studies the prediction of newborn Length of Stay. The topic is relevant and useful for hospital management. The paper has good potential, but it needs major revisions before it can be accepted.

The main issue is the research gap, the novelty is not clearly written, and the formatting.

has the title written properly?

Yes, the authors has mentioned method used in the title

has the abstract section written properly?

Yes, the authors mention the impact/urgency/importance of this research for informatics/computer science?

has the introduction section written properly?

No, Authors should mention/add a gap that becomes the novelty/originality of this research compared to other research In the introduction.

In the introduction, are there a minimum of 25 references?

Yes, there are minimum of 25 references in introduction

has the methods section written properly?

Yes, the methods chapter has provided information on how to collect data, how to process data, and an explanation of the evaluation carried out

has the result section written properly?

Yes, the results section has been arranged according to the sequence of stages in the methods chapter.

Has each table and figure been referenced and explained by including the figure number and table number?

Yes, each table and figure has been referenced and explained by including the figure number and table number

has the discussion section written properly?

No, the authors should add/mention comparison the results of the research conducted by referring to other research/journals that have the same topic/same data and convey the impact/urgency/importance of this research for informatics/computer science

has the conclusion section written properly?

No, the authors should add/mention and conveyed the impact/urgency/importance of this research on the science of Informatics/Computer Science

has the references section written properly?

- The bibliography come from the journal has not yet completed with a DOI.
- Bibliography come from the Journal has not yet completed with journal name, volume, number, and page.

-----

-----  
Reviewer B:

Recommendation: Revisions Required

-----

Title

Judul sudah sesuai, yaitu sudah mengandung objek penelitian, metode dan tujuan penelitian.  
Tidak menggunakan singkatan atau istilah umum yang tidak dijelaskan.

Abstract

Abstrak sudah memuat masalah pada penelitian, tujuan, metode dan hasil penelitian.  
Saran penulisan dampak atau urgensi penelitian agar lebih eksplisit.

Kata kunci sudah berisi 5 frase yang berkaitan dengan isi penelitian dan disusun secara alphabet.

## Introduction

Pendahuluan sudah menyajikan masalah atau pentingnya penelitian.  
Mohon tambahkan gap dari penelitian yang dibandingkan dengan penelitian lain secara jelas.  
Rujukan pada pendahuluan minimal 25, maks 5 tahun terakhir.  
Perlu penambahan penegasan novelty di akhir sebelum tujuan.  
Tujuan belum ditulis eksplisit di akhir paragraf.

## Method

Metode belum menampilkan tahapan penelitian menggunakan gambar.  
Setiap gambar dan tabel diberikan penjelasan dan disebut penomorannya dengan benar pada teks.  
Penulisan equation mohon cantumkan angka penomorannya di samping kanan rumus, dan dirujuk pada teks. Sesuaikan dengan template Jutif.  
Mohon penjelasan proses feature engineering (ICD-10 grouping) dapat diperjelas.  
Perlu penjelasan tipe data (numeric/categorical) + total fitur model.

## Results

Penulisan hasil sebaiknya diurutkan berdasarkan urutan tahapan pada bagian Metode, dan harus memuat hasil secara keseluruhan dari penelitian.  
Penomoran gambar dan tabel harus dirujuk pada teks.  
Mohon rapikan Tabel 1.  
Tambahkan tabel "Feature Importance" untuk Random Forest.

## Discussions

Pada bagian Discussion sudah dituliskan pembahasan sesuai pendapat penulis mengenai hasil penelitiannya yang telah dilakukan secara detail.  
Tambahkan perbandingan dengan penelitian sebelumnya.  
Tambahkan dampak atau urgensi dari penelitian ini dengan lebih jelas.

## Conclusion

Bab kesimpulan sudah merangkum isi dari hasil diskusi dan pembahasan dan terdapat saran penelitian lanjutan.  
Tambahkan kalimat eksplisit tentang kontribusi penelitian ke ilmu komputer/informatika.

## References

Jumlah lebih dari 25 rujukan 80% dari jurnal/conference terbitan 5 tahun terakhir dan sudah menggunakan style IEEE.  
Mohon perbaiki metadata pada Mendeley atau Zotero, atau EndNote, ada beberapa referensi DOI belum lengkap, nama jurnal tidak lengkap.  
Saran untuk referensi yang terbit lebih dari 5 tahun mohon dapat diupdate dengan referensi yang terbaru.

## Overall Comment

Revisi Minor

has the title written properly?

Yes, the authors has mentioned method used in the title

has the abstract section written properly?

Yes, the authors mention the impact/urgency/importance of this research for informatics/computer science?

has the introduction section written properly?

No, Authors should mention/add a gap that becomes the novelty/originality of this research compared to other research In the introduction.

In the introduction, are there a minimum of 25 references?

No, the authors should add citation minimum 25 references in introduction

has the methods section written properly?

Yes, the methods chapter has provided information on how to collect data, how to process data, and an explanation of the evaluation carried out

has the result section written properly?

No, the results section should be arranged according to the sequence of stages in the methods chapter.

Has each table and figure been referenced and explained by including the figure number and table number?

No, the authors should add/mention each table and figure to be been referenced and explained by including the figure number and table number

has the discussion section written properly?

No, the authors should add/mention comparison the results of the research conducted by referring to other research/journals that have the same topic/same data and convey the impact/urgency/importance of this research for informatics/computer science

has the conclusion section written properly?

No, the authors should add/mention and conveyed the impact/urgency/importance of this research on the science of Informatics/Computer Science

has the references section written properly?

- References section has written properly and completed

-----

---

Jurnal Teknik Informatika (Jutif)

---

## 2 lampiran



**A-jutif-review-assignment-5410-Article+Manuscript-16898 - Reviewed.docx**

168K



**B-jutif-review-assignment-5410-Article+Manuscript-16898\_rev.docx**

171K

[jutif] Editor Decision

2025-12-04 06:37 AM

Gandung Triyono:

We have reached a decision regarding your submission to Jurnal Teknik Informatika (Jutif), with e-ISSN: 2723-3871 and is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024, with title "Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay".

Our decision is: Revisions Required

In addition, several things to consider from the editor's side are as follows :

- 1. Please **FOLLOW and USE** JUTIF's template. You can download it at <http://jutif.if.unsoed.ac.id/index.php/jurnal/about/submissions>. After that, you can copy your paper into JUTIF's template.
- 2. Please use table and figure format according to JUTIF's Template. The resolution of figures must be clear, can be seen and **more than 300 dpi**.
- 3. Every figure and table **MUST BE** cited and must be explained in paragraph, ex : "Research method can be seen in Figure 1", "Table 1 showed XYZ."
- 4. Section "Discussion" is **MANDATORY**. If your article did not contain Discussion, please add it.
- 5. **MINIMUM** references in JUTIF is **25 primary references** (Journal/Conference) from 5 years latest. Please add reference if necessary. In JUTIF, Reference using IEEE format. Please adjust in your reference. We recommend to use reference manager tools, such as : Mendeley, Zootero, or EndNote.
- 6. Please Correct the manuscript **according to the review** from the reviewer or editor. Comments from reviewers that need to be corrected for the manuscript has been moved at the bottom.
- 7. Please **give mark the revised results on the revised manuscript sent with YELLOW HIGHLIGHTS** to the corrected sentence or sentences added of revision fulfillment from the Reviewer.
- 8. Please upload revision file along with Turnitin Result File (If Any). The **Maximum plagiarism result is 20%**.
- 9. Revisions can be sent via the "File upload" in the "revision" section of the JUTIF's OJS. Please when upload the Revised File, do not overwrite the previous revised file (if any).
- 10. For Information, you can contact our representative : 0823-2492-4093.

Thank You.

Best Regards.

Reviewer A:  
Recommendation: Revisions Required

Title  
  
Novelty not reflected in the title.  
Could add dataset context, domain, etc.

-----  
Reviewer A:  
Recommendation: Revisions Required  
-----

#### Title

Novelty not reflected in the title.  
Could add dataset context, domain, etc.

#### Abstract

Could add explicit gap statements and novel contributions.

#### Introduction

Could add a paragraph to explain the research gap.  
The literature review is descriptive; it should be more comparative.

#### Method

- Could add a figure for the research framework, showing:  
Data Collection → Preprocessing → Feature Engineering → Modeling → Evaluation
- Could add a table of dataset attributes.
- Fix equations formatting. Read: Jutif Template.
- In Methods, should focus on implementation, not just on mathematical algorithms.

#### Results

- Add more descriptive results.
  - The table format should follow Jutif style. Read: Jutif template.
- Download Reviewer Files for more details.

#### Discussions

- Add comparison with the previous study.
- Add more explanation of why RFR works better for this dataset.
- Add practical implications for Hospitals.
- Could discuss the limitations.

#### Conclusion

Add an explicit contribution statement.

#### References

- Some DOIs are missing and the journal names are not in full format. Fix it.
- The WHO website citation should not dominate.
- Correct spelling errors in the references.
- Preprints references should be minimized.

#### Overall Comment

This manuscript studies the prediction of newborn Length of Stay. The topic is relevant and useful for hospital management. The paper has good potential, but it needs major revisions

before it can be accepted.

The main issue is the research gap, the novelty is not clearly written, and the formatting.

has the title written properly?

Yes, the authors has mentioned method used in the title

has the abstract section written properly?

Yes, the authors mention the impact/urgency/importance of this research for informatics/computer science?

has the introduction section written properly?

No, Authors should mention/add a gap that becomes the novelty/originality of this research compared to other research In the introduction.

In the introduction, are there a minimum of 25 references?

Yes, there are minimum of 25 references in introduction

has the methods section written properly?

Yes, the methods chapter has provided information on how to collect data, how to process data, and an explanation of the evaluation carried out

has the result section written properly?

Yes, the results section has been arranged according to the sequence of stages in the methods chapter.

Has each table and figure been referenced and explained by including the figure number and table number?

Yes, each table and figure has been referenced and explained by including the figure number and table number

has the discussion section written properly?

No, the authors should add/mention comparison the results of the research conducted by referring to other research/journals that have the same topic/same data and convey the impact/urgency/importance of this research for informatics/computer science

has the conclusion section written properly?

No, the authors should add/mention and conveyed the impact/urgency/importance of this research on the science of Informatics/Computer Science

has the references section written properly?

- The bibliography come from the journal has not yet completed with a DOI.
- Bibliography come from the Journal has not yet completed with journal name, volume, number, and page.

-----  
Reviewer B:

Recommendation: Revisions Required  
-----

Title

Judul sudah sesuai, yaitu sudah mengandung objek penelitian, metode dan tujuan penelitian. Tidak menggunakan singkatan atau istilah umum yang tidak dijelaskan.

Abstract

Abstrak sudah memuat masalah pada penelitian, tujuan, metode dan hasil penelitian. Saran penulisan dampak atau urgensi penelitian agar lebih eksplisit. Kata kunci sudah berisi 5 frase yang berkaitan dengan isi penelitian dan disusun secara alphabet.

Introduction

Pendahuluan sudah menyajikan masalah atau pentingnya penelitian. Mohon tambahkan gap dari penelitian yang dibandingkan dengan penelitian lain secara jelas. Rujukan pada pendahuluan minimal 25, maks 5 tahun terakhir. Perlu penambahan penegasan novelty di akhir sebelum tujuan. Tujuan belum ditulis eksplisit di akhir paragraf.

Method

Metode belum menampilkan tahapan penelitian menggunakan gambar. Setiap gambar dan tabel diberikan penjelasan dan disebut penomorannya dengan benar pada teks. Penulisan equation mohon cantumkan angka penomorannya di samping kanan rumus, dan dirujuk pada teks. Sesuaikan dengan template Jutif. Mohon penjelasan proses feature engineering (ICD-10 grouping) dapat diperjelas. Perlu penjelasan tipe data (numeric/categorical) + total fitur model.

Results

Penulisan hasil sebaiknya diurutkan berdasarkan urutan tahapan pada bagian Metode, dan harus memuat hasil secara keseluruhan dari penelitian.

Penomoran gambar dan tabel harus dirujuk pada teks.

Mohon rapikan Tabel 1.

Tambahkan tabel "Feature Importance" untuk Random Forest.

#### Discussions

Pada bagian Discussion sudah dituliskan pembahasan sesuai pendapat penulis mengenai hasil penelitiannya yang telah dilakukan secara detail.

Tambahkan perbandingan dengan penelitian sebelumnya.

Tambahkan dampak atau urgensi dari penelitian ini dengan lebih jelas.

#### Conclusion

Bab kesimpulan sudah merangkum isi dari hasil diskusi dan pembahasan dan terdapat saran penelitian lanjutan.

Tambahkan kalimat eksplisit tentang kontribusi penelitian ke ilmu komputer/informatika.

#### References

Jumlah lebih dari 25 rujukan 80% dari jurnal/conference terbitan 5 tahun terakhir dan sudah menggunakan style IEEE.

Mohon perbaiki metadata pada Mendeley atau Zotero, atau EndNote, ada beberapa referensi DOI belum lengkap, nama jurnal tidak lengkap.

Saran untuk referensi yang terbit lebih dari 5 tahun mohon dapat diupdate dengan referensi yang terbaru.

#### Overall Comment

##### Revisi Minor

has the title written properly?

Yes, the authors has mentioned method used in the title

has the abstract section written properly?

Yes, the authors mention the impact/urgency/importance of this research for informatics/computer science?

has the introduction section written properly?

No, Authors should mention/add a gap that becomes the novelty/originality of this research compared to other research In the introduction.

In the introduction, are there a minimum of 25 references?

No, the authors should add citation minimum 25 references in introduction

has the methods section written properly?

Yes, the methods chapter has provided information on how to collect data, how to process data, and an explanation of the evaluation carried out

has the result section written properly?

No, the results section should be arranged according to the sequence of stages in the methods chapter.

Has each table and figure been referenced and explained by including the figure number and table number?

No, the authors should add/mention each table and figure to be been referenced and explained by including the figure number and table number

has the discussion section written properly?

No, the authors should add/mention comparison the results of the research conducted by referring to other research/journals that have the same topic/same data and convey the impact/urgency/importance of this research for informatics/computer science

has the conclusion section written properly?

No, the authors should add/mention and conveyed the impact/urgency/importance of this research on the science of Informatics/Computer Science

has the references section written properly?

- References section has written properly and completed

-----

6. Hasil Review dari Reviewer 1 pada File

Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025  
Phone Number : xxxxxxx (Whatsapp/Cellphone Number, For Corresponding)  
The cellphone number is only for ease of communication and will NOT be displayed in the article

Abstract

The length of stay (LOS) of newborns is a crucial indicator for healthcare management and hospital resource allocation. However, the factors influencing LOS are not yet fully understood, and existing prediction methods often yield suboptimal results. This study aims to identify the factors affecting LOS and to develop a predictive model for newborn LOS using several machine learning algorithms. A comparison was conducted among Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN) to determine the most effective model. The data were obtained from the medical records of newborn patients at three private hospitals in Indonesia. The research followed several stages, including data collection and understanding, data preprocessing, modeling, and evaluation. The experimental results demonstrated that the Random Forest Regression algorithm achieved the best predictive performance, with MAE = 0.019, MSE = 0.011, RMSE = 0.086, and R<sup>2</sup> = 0.987. The feature importance analysis of the Random Forest model revealed that gender, referral source, insurance type, and diagnosis were the most influential factors affecting LOS. Theoretically, this study strengthens the application of machine learning in healthcare data analytics, particularly in predicting the length of stay for newborn patients.

Keywords : length of stay, machine learning, newborns, prediction, Random Forest;

This work is an open access article licensed under a Creative Commons Attribution 4.0 International License.



1. INTRODUCTION

Newborn health serves as a key indicator for assessing the quality of a country's healthcare services. Infant mortality and morbidity rates reflect not only the health status of mothers and children but also the overall effectiveness of the healthcare system [1]. As reported by Statistics Indonesia in 2020, although Indonesia has experienced a decline in infant mortality over the past few decades, regional disparities remain significant, indicating inequality in access to and quality of healthcare services. Certain regions, such as Maluku and Papua, continue to record higher infant mortality rates compared to the national average, while areas like Java are approaching the Sustainable Development Goals (SDGs) target of 12 deaths per 1,000 live births.

Accurately predicting the LOS of newborns holds significant implications for hospitals in terms of resource management and cost efficiency[2], [3]. Reliable prediction models enable healthcare providers to design more effective care strategies, optimize hospital bed utilization, and reduce patients' financial burdens [3], [4]. However, traditional approaches such as logistic regression and survival analysis often fail to adequately address the non-linear and complex nature of medical data [5], [6].

One of the critical aspects of neonatal care is the length of hospital stay (LOS), which varies considerably depending on several medical factors such as prematurity, respiratory disorders, infections, and congenital anomalies [7]. Prematurity, for example, is a major contributor to extended hospitalization due to the incomplete development of vital organs, particularly the lungs and immune system [3]. Similarly, Respiratory Distress Syndrome (RDS) is frequently cited as a primary reason for prolonged admission in neonatal intensive care units [8]. In addition, LOS is also influenced by age, gender, and social determinants [4].

Commented [A1]: Novelty not reflected in the title. Could add dataset context, domain, etc.

Commented [A2]: Could add explicit gap statements and novel contribution.

Commented [A3]: Could add a paragraph to explain research gap. The literature review is descriptive, should be more comparative.

With the advancement of data-driven technologies, machine learning (ML) has emerged as a promising alternative for predictive analytics in healthcare. Numerous studies have demonstrated its potential in forecasting hospital stay durations. Almeida [3] reported that the Random Forest algorithm can produce highly accurate predictions when optimal feature selection is applied. Barnes [9] utilized real-time data from pediatric intensive care units. Moreover, Zeng [10] found that LightGBM achieved superior predictive performance, and Shbool et al. [6], [11] demonstrated that Random Forest can effectively enhance hospital resource utilization, particularly during the COVID-19 pandemic.

In this study, several machine learning algorithms—Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—are employed to predict neonatal LOS. These algorithms were selected for their respective strengths in handling data with varying levels of complexity. LR serves as the baseline model to capture linear relationships between input and output variables. RFR is chosen for its robustness in modeling non-linear interactions and combining multiple decision trees to enhance predictive accuracy [6], [11], [12]. SVR offers flexibility in capturing non-linear patterns through the use of kernel functions, while ANN is utilized for its capability to learn complex and high-dimensional representations in medical datasets, thereby improving predictive precision [13].

Model performance is evaluated using four quantitative metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared ( $R^2$ ). MAE measures the average magnitude of errors in the same units as the data, MSE penalizes larger deviations more heavily, RMSE retains interpretability while accounting for outliers, and  $R^2$  assesses the proportion of variance explained by the model. Collectively, these metrics provide a comprehensive framework for assessing model performance [10], [11].

The primary objective of this study is to identify key determinants influencing the LOS of newborns and to develop more accurate predictive models compared to conventional statistical methods. Understanding these determinants is crucial for hospitals to optimize resource allocation—including bed capacity, medical personnel, and neonatal care equipment. The analysis incorporates both medical and non-medical factors, such as gender, clinical conditions (e.g., prematurity, respiratory distress, or infection), and socio-economic characteristics that may affect hospitalization outcomes.

Furthermore, this study seeks to compare the predictive performance of different machine learning models in capturing the non-linear structures inherent in medical data. The evaluation using MAE, MSE, RMSE, and  $R^2$  provides a detailed comparison of the predictive reliability and explanatory power of each algorithm. Overall, this research not only advances predictive modeling in healthcare but also supports the broader goal of improving the quality of care for newborn patients in hospital settings.

## 2. METHOD

To ensure consistency throughout the research process, this study was structured into several stages that align with its objectives. The main phases of the research include data collection, data preprocessing, modeling, and model evaluation. Each of these stages is described in detail in the following subsections.

### 2.1. Data Collection

The first step of this research involves understanding the business requirements and primary objectives of the study. Hospitals often face challenges in managing bed capacity and medical resources, particularly for newborn patients who require specialized care. Therefore, this study aims to develop a predictive model to estimate the length of stay (LOS) based on both medical and non-medical factors. By utilizing such a model, hospitals can plan resource allocation more effectively and enhance service efficiency.

**Commented [A4]:** Could add figure for research framework.

Could add table of dataset attributes.

Fix equations formatting. Read: Jutif Template.

In the Method, could focus on implementation.

The study uses data from newborn patients hospitalized in selected hospitals, encompassing clinical and demographic variables such as infant initial diagnosis, medical interventions, and other relevant medical and non-medical factors influencing the duration of hospitalization.

1. Population: All newborns admitted to three hospitals during the study period.
2. Inclusion Criteria: Newborns with complete medical records who were hospitalized during the year 2023.
3. Exclusion Criteria: Newborns with incomplete medical records or those referred to other hospitals before completing treatment.

## 2.2. Data Preprocessing

This stage involves several data preprocessing steps to ensure the dataset is in optimal condition before being used for machine learning modeling:

- Data Deduplication: Removing duplicate entries to maintain the accuracy and integrity of the dataset.
- Data Cleansing: Handling missing values using imputation methods (mean or median) and removing inconsistent or extreme data points.
- Data Transformation: Applying normalization to numerical features and encoding categorical features to ensure the data is properly formatted for model training.
- Dimensionality Reduction: Reducing the dimensionality of diagnostic features by utilizing ICD-10 codes, extracting the corresponding *Block* and/or *Chapter* levels for feature representation.
- Data Splitting: Dividing the dataset into a training set (80%) and a testing set (20%) to ensure fair model training and evaluation. Additionally, cross-validation techniques are employed to improve model generalization.

## 2.3. Modeling

Based on the studies that have been conducted, this study uses four machine learning algorithms to build a prediction model for the length of hospitalization, namely LR, RFR, SVR and ANN.

### a. Linear Regression

As a supervised machine learning technique, linear regression is trained on labeled datasets to construct the best-fit linear function. This optimized function aims to capture the linear relationship the idea that the output shifts uniformly (at a constant rate) whenever the input changes and is subsequently employed for forecasting values on unseen data. The relationship itself is fundamentally depicted by a straight line [14], [15]. The model of linear regression can be expressed in formula (1).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

where Y represents the dependent variable, which corresponds to the predicted output, while X denotes the independent variable or predictor.  $\beta_0$  refers to the intercept (constant term),  $\beta_1$  is the regression coefficient associated with X, and  $\varepsilon$  represents the error term.

In simple linear regression, the regression coefficients  $\beta_0$  and  $\beta_1$  can be calculated using the Ordinary Least Squares (OLS) method with the formula:

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad (3)$$

where  $X_i$  and  $Y_i$  represent the dependent and independent variables, respectively.  $\bar{X}$  denotes the mean value of X,  $\bar{Y}$  denotes the mean value of Y, and n refers to the total number of data points.

If there is more than one independent variable, then the model becomes multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (4)$$

where  $Y$  is the dependent variable, namely the predicted output.  $\beta_0$  is an intercept or constant.  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficient for each independent variable.  $X_1, X_2, \dots, X_n$  are an independent variable, namely the predictor.  $\varepsilon$  is an error.

The following are the steps in the modeling process using a linear regression algorithm in a study on predicting the length of hospitalization for newborns:

- Determine the input variables ( $X_1 \dots X_n$ ), such as gender, initial diagnosis, and insurance, which are suspected of influencing  $Y$  (length of hospitalization).
- Standardize or rescale the data to avoid bias due to differences in units between features before the regression process.
- Apply a linear model to map the relationship between  $X_1, \dots, X_n$  and  $Y$  using the OLS approach.
- Calculate the values of  $\beta_0, \beta_1, \dots, \beta_n$  that minimize the residual error  $\varepsilon$ , so that the regression line approximates the actual patient length of stay.

#### b. Random Forest Regression

Random Forest Regression (RFR) is a powerful ensemble machine learning technique that utilizes a collection of decision trees to achieve a more robust and accurate prediction. The central goal of RFR is to prevent the high variance and overfitting issues typical of single decision trees [16].

It achieves this by injecting randomness during the training process:

- Each tree is built using a bootstrap sample (a random subset with replacement) of the original data.
- At every node split, only a random subset of features is considered.

By combining and averaging the predictions from these diverse, independently-built trees, the final RFR model improves generalization and stability. Furthermore, Out-of-Bag (OOB) samples—the data excluded from each tree's training set—are used for internal performance validation.

The use of RFR learning ensembles, capable of producing models with low variability and higher accuracy. Eq. (3) represent RFR:

$$R(a) = \sum_{k=1}^N Y_k \sum_{j \in C_k} \theta(j; p_k) = \sum_{k=1}^N \sum_{j \in C_k} Y_k \theta(j; p_k) \quad (5)$$

with  $R(a)$  is the final decision of the classifier,  $Y_k$  average response value ke-k,  $C_k, P_k$  is a divisive parameter to divide the region of the decision and  $\theta(j; p_k)$  is a function of the information limit based on the  $P_k$  and  $j$ -index. This equation shows the aggregation of the response values in each region  $k$ , taking into account the data subes and relevant divisor parameters [17].

#### c. Support Vector Regression

Support Vector Regression (SVR) is an extension of SVM designed to address regression problems. In SVM, the primary goal is to find the best hyperplane that separates data classes. Conversely, in SVR, the primary goal is to find a function that predicts continuous values based on the input, while maintaining the SVM's margin principles. SVR works by building a model that not only accounts for prediction errors but also minimizes model complexity. To do this, SVR uses an epsilon-insensitive loss function, which allows small errors to be ignored if they are within an epsilon distance from the predicted value [18].

The Support Vector Regression (SVR) modeling process consists of several key steps. First, a mathematical model is constructed to map input variables, such as diagnosis and gender, to the target variable, namely the length of stay. A tolerance margin is then established to define the acceptable range of prediction errors that do not incur penalties. Errors falling within this range are disregarded to reduce overfitting, while those exceeding the tolerance threshold receive larger penalties. The model is subsequently optimized to maintain a balance between simplicity and predictive accuracy. In cases where the relationships between variables are non-linear, kernel functions are applied to transform the data into a higher-dimensional space, thereby facilitating better pattern recognition. Support vectors—data points located on or beyond the tolerance boundaries—are then identified, as they play a crucial role in determining the structure of the regression function. Finally, the trained SVR model is used to predict the length of stay for new patient data.

SVR maps a low-dimensional input space into a high-dimensional feature space using a nonlinear mapping and then performs a linear regression in the high-dimensional feature space. Given the input dataset  $D = \{(x_i, y_i) : i = 1 \cdots N\}$ , the wind power prediction of SVR method can be given as  $y = f(x) = \omega \phi(x) + bx$ , where  $w$  is the weight,  $b$  is the parameter of bias, and  $\phi(x)$  represents the high-dimensional space. To obtain parameters  $w$  and  $b$ , the prediction of SVR model with structure risk minimization principle can be transformed to a mathematical optimization problem and solved using Lagrange multiplier [19]. The optimization problem is presented as follows:

$$f(X, Y_i, Y_i^*) = \sum_{i=1}^N (Y_i - Y_i^*) K(X, X_i) + b \quad (6)$$

where  $Y_i$  and  $Y_i^*$  are the Lagrange multiplier and  $K(X, X_i)$  represents the kernel function. In this study, the radial-basis function is adopted as the kernel function of SVR. Some important parameters of SVR-like kernel coefficient, penalty parameter of the error term and radius are selected by many trails with empirical data.

#### d. Artificial Neural Network

An Artificial Neural Network (ANN) is a computer model inspired by how the human brain processes information. It is made up of layers of connected neurons, which allow the network to learn more complex patterns in data as it trains. ANNs work by transforming input data through multiple layers, with each layer learning to extract more meaningful features. The network is trained using backpropagation, a process that adjusts the connection weights based on the difference between the network's output and the expected target [20].

A common ANN structure is the multi-layer backpropagation network, which has three types of layers: the input layer, hidden layers, and output layer. The input layer receives the input variables ( $X_1, X_2, \dots, X_n$ ). The hidden layers process this information ( $Z_1, Z_2, \dots, Z_n$ ), while the output layer produces the final results ( $Y$ ), such as sample classifications.

The network updates the weights of each layer to reduce errors. The error shows how well the network is learning—it is calculated as the difference between the network's current output and the desired output.

During forward propagation, the network calculates the weighted sum of inputs for each neuron. If the input signal passes a certain threshold, it goes through an activation function, which determines the neuron's output. Activation functions must be continuous, differentiable, and non-decreasing to work with backpropagation. A common example is the Binary Sigmoid function (as shown in Equations (7) and (8)).

$$Y = f(X) = \frac{1}{1 + e^{-\sigma x}} \quad (7)$$

$$f(X) = \sigma f(X)[1 - f(X)] \quad (8)$$

The following explains how the multilayer perceptron model operates using backpropagation.

Step 0: Initialize the weights, preferably set to small random values,

Step 1: If the condition is not met, do steps 2-9,

Step 2: For each training pair, perform steps 3-8.

Steps 3-5 are called forward propagation.

Step 3: Each input unit ( $X_i, i = 1, 2, \dots, n$ ) receives the  $X_i$  signal and transmits this signal to all units in the layer above it (hidden units),

Step 4: Each hidden unit ( $X_i, i = 1, 2, \dots, p$ ) sums the weights of its input signals,

$$Z\_in_j = V_{oj} + \sum_{i=1}^n X_i V_{ij} \quad (9)$$

$V_{oj}$  = bias on hidden unit  $j$  apply its activation function to calculate its output signal,

$Z_j = f(Z\_in_j)$ , and send this signal to all units in the layer above it (output units).

Step 5: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ), summing the weights of its input signals,

$$Y\_in_k = W_{ok} + \sum_{j=1}^n Z_j W_{jk} \quad (10)$$

$W_{ok}$  = bias on the output unit  $k$  and apply its activation function to calculate its output signal,  $Y_k = f(Y\_in_k)$ .

In Steps 6-8 it is called backpropagation:

Step 6: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ) receive the interconnected target patterns at the input of the training patterns, calculate the information error,

$$\Delta_k = (T_k - Y_k) f^1(T\_in_k) \quad (11)$$

calculate the weight correction (used to update the  $W_{jk}$  later),

$$\Delta W_{jk} = \alpha \Delta_k Z_j \quad (12)$$

calculate the bias correction (used to update the  $W_{ok}$  later), and send  $\delta_k$  to the units in the layer below,

Step 7: Each layer unit is hidden ( $Z_j, j = 1, 2, \dots, p$ ) sum the results of the input changes (from the units of the layers above),

$$\Delta\_in_j = \sum_{k=1}^n \Delta_k W_{jk} \quad (13)$$

multiply by the derivative of its activation function to calculate its error information,

$$\Delta_j = \Delta\_in_j f^1(Z\_in_j) \quad (14)$$

calculate the weight correction (used to update  $V_{oj}$  later),

Step 8: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ) update bias and weights ( $j = 0, 1, \dots, p$ ):

$$W_{jk}(\text{new}) = W_{jk}(\text{old}) + \Delta W_{jk} \quad (15)$$

Each hidden layer unit ( $Z_j, j = 1, 2, \dots, p$ ) update bias and weights ( $i = 0, 1, 2, \dots, n$ ):

$$V_{ij}(\text{new}) = V_{ij}(\text{old}) + \Delta V_{ij} \quad (16)$$

Step 9: Test the stopping condition.

Steps of the testing process:

Following the training phase, the backpropagation neural network is applied strictly using its forward propagation stage. The application steps are as follows:

Step 0: Initialize weights (from the training algorithm).

Step 1: for each input vector, perform the steps 2-4.

Step 2: for  $i = 1, 2, 3, \dots, n$ : set the activation of the  $X_i$  input unit.

Step 3: for  $j = 1, 2, \dots, p$ :

$$Z\_in_j = V_{oj} + \sum_{i=1}^n X_i V_{ij} \quad (17)$$

$$Z_j = f(Z\_in_j) \quad (18)$$

Step 4: for  $k = 1, 2, \dots, m$ :

$$Y\_in_k = V_{ok} + \sum_{j=1}^n X_j V_{jk} \quad (19)$$

$$Y_k = f(Z\_in_k) \quad (20)$$

Step 5: if  $Y_k \geq 0,5$  then  $Y_k = 1$ , else  $Y_k = 0$ .

## 2.4 Evaluation Model

After modeling, the next step is to test the model's performance using various evaluation metrics, namely MSE, RMSE, MAE, and  $R^2$ . This testing is done to determine which model provides the most accurate predictions.

Mean Squared Error (MSE) is an equation that measures the average squared error between actual and predicted values. A smaller MSE value indicates better model performance (predictions are closer to reality) [14], [21], as expressed in the following equation:

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (21)$$

where  $n$  merupakan sample datam  $y_i$  represents the true value,  $\hat{y}_i$  denotes the predicted value, and  $(y_i - \hat{y}_i)^2$  denotes the squared difference between the actual and predicted values.

The Root Mean Squared Error (RMSE) is a widely used statistical metric that measures the average magnitude of the prediction errors in a regression model. It is defined as the square root of the mean of the squared differences between the actual and predicted values [14], [22], as expressed in the following equation:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (22)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Mean Absolute Error (MAE) is a commonly used evaluation metric that measures the average magnitude of the errors in a set of predictions, without considering their direction. It is defined as the mean of the absolute differences between the actual and predicted values [14], [22], as expressed in the following equation:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (23)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Coefficient of Determination ( $R^2$ ), commonly referred to as R-squared, is a statistical metric used to evaluate the goodness of fit of a regression model. It represents the proportion of the variance in the actual (observed) data that is explained by the predicted values generated by the model [23], [24]. The mathematical formulation of  $R^2$  is expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (24)$$

where  $y_i$  denotes the actual value,  $\hat{y}_i$  represents the predicted value,  $\bar{y}$  is the mean of the actual values, and  $n$  is the total number of observations. The numerator  $\sum (y_i - \hat{y}_i)^2$  represents the Residual Sum of Squares (RSS), while the denominator  $\sum (y_i - \bar{y})^2$  represents the Total Sum of Squares (TSS).

An  $R^2$  value close to 1 indicates that the model explains most of the variability in the data, signifying a strong predictive capability. Conversely, an  $R^2$  value near 0 suggests that the model fails to capture the underlying data pattern. In some cases,  $R^2$  can be negative if the model performs worse than a simple mean-based prediction.

3. **RESULT**

The study results are organized into three subsections: Data Collection and Preprocessing, Modeling, and Model Evaluation. The first subsection details the dataset and preprocessing outcomes, the second compares the RL, RFR, SVR, and ANN algorithms, and the third presents the evaluation results.

3.1. **Data Collection and Preprocessing**

The length of stay (LOS) for newborn infants plays a vital role in planning medical care and managing hospital resource allocation. This study aims to determine the factors influencing neonatal LOS and to develop a machine learning-based prediction model that enhances efficiency and accuracy in hospital management. The data used in this research were collected from three private hospitals in Indonesia.

The dataset successfully collected comprised 3.244 cases obtained from medical records. The data included inpatient information with attributes such as gender, type of insurance, diagnosis, referral procedure, referral type, and length of stay (LOS). Following data cleaning and validation in collaboration with hospital physicians, a refined dataset containing 445 cases was produced. This cleaned dataset was subsequently utilized for model development. Table 1 provides an illustration of the dataset employed in predicting the length of stay (LOS) of newborn patients.

Table 1. Example of a Dataset Used for Predicting the Length of Stay (LOS) of Newborns

| No  | Gender | Diagnosis | Insurance Claim | Referral Procedure                          | Referral Patient | LOS |
|-----|--------|-----------|-----------------|---------------------------------------------|------------------|-----|
| 1   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 2   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 3   | Male   | P22.9     | Government      | Umbilical venous catheter                   | Yes              | 2   |
| 4   | Female | P07.2     | Private         | Parenteral nutrition                        | Yes              | 1   |
| 5   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 5   |
| ... | ...    | ...       | ...             | ...                                         | ...              | ... |
| 6   | Male   | Z38.0     | Private         | Single liveborn infant, delivered vaginally | No               | 1   |

Gender are crucial demographic factors that significantly influence how severe a disease becomes. Generally, the risk of a severe illness escalates with advancing age, particularly when patients have comorbidities and declining organ function. This increase in risk directly impacts the patient's Length of Stay (LOS) in the hospital[7], [25], [26]. The collected data indicated that male infants constituted 236 cases (53%), whereas female infants accounted for 209 cases (47%).

Insurance claim data has been utilized in several studies to predict LOS; however, only a limited number of investigations have employed hospital records for LOS prediction [13]. In this study, the insurance attribute represents the type of coverage held by the patient, categorized as government insurance, private insurance, or self-payment [27]. Based on the data utilized in this study, 224 cases (50.3%) were covered by government insurance, 218 cases (49.0%) were covered by private insurance, and only 3 cases (0.7%) were paid out-of-pocket.

Diagnosis refers to the determination of a disease, condition, or injury based on observed signs and reported symptoms. It involves gathering and analyzing patient history, conducting physical examinations, and interpreting medical test results to identify the underlying cause and inform appropriate treatment and prognosis. Beyond the medical field, the concept of diagnosis is also applied in areas such as science, engineering, and business to investigate issues and pinpoint their root causes [2].

Commented [A5]: Add more descriptive results.

Commented [A6]: The table format should follow Jutif style.

A referral procedure is a series of steps for transferring a patient from one healthcare facility to another in order to obtain more specialized treatment. Typically, this process begins at a Primary Healthcare Facility and is continued to a higher-level Referral Healthcare Facility when medically necessary. Patients must first undergo an examination at the primary facility, where a physician will decide whether a referral is required. However, in emergency situations, patients may go directly to the nearest hospital without a referral.

Healthcare referral pathways are generally categorized into medical referrals and general health referrals. In the dataset utilized for this study, referral type is represented by two classifications: health referrals and emergency referrals. Of the 445 recorded inpatient cases involving newborns, only 10 cases were identified as emergency referrals, while 1 case involved a newborn admitted in stable, normal condition. Overall, these data indicate that approximately 2.5% of newborn hospitalization cases originated from referral pathways.

Based on Table 1, LOS refers to the duration of a patient's hospitalization. In this study, LOS is categorized into six groups: Group 0 (0-day stay), Group 1 (1–7 days), Group 2 (8–14 days), Group 3 (15–21 days), Group 4 (22–30 days), and Group 5 (more than 30 days). Figure 3 illustrates the distribution of these length-of-stay categories. The results of the analysis can be explained that the LOS data is not balanced from 6 groups, 66.3% for group 1 (see Figure 1).

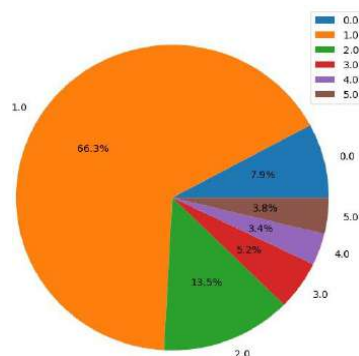


Figure 1. Distribution of LOS Data

As illustrated in Figure 2, the distribution of diagnoses among newborn inpatients includes 145 cases related to immunization-associated health issues, 97 cases involving respiratory and metabolic disorders, 82 cases classified under post-procedural and treatment complications, and 23 cases associated with prematurity. A detailed breakdown of this distribution is presented in Figure 2.

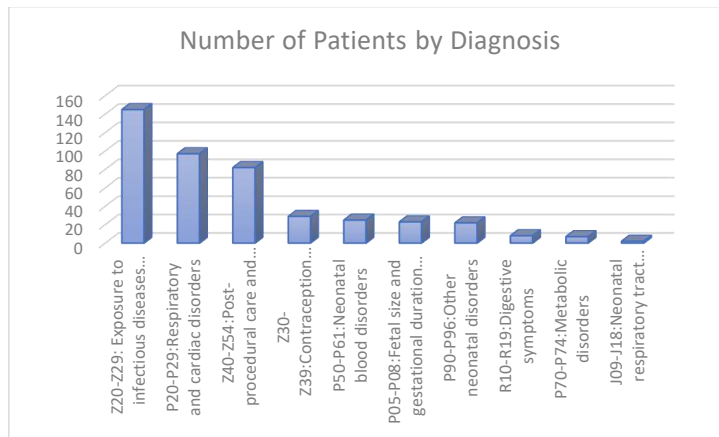


Figure 2. Data Distribution for Diagnosis

### 3.2. Modeling

At this stage, models for regression are developed and optimized using four different algorithms: LR, RFR, SVR, and Artificial Neural Network (ANN), each employing GridSearchCV to identify the best parameters.

- Linear Regression Pipeline: Data preprocessing (numeric features unchanged, categorical features OneHotEncoded) is followed by a Linear Regression model, with GridSearchCV (5-fold CV) used to find the best hyperparameters based on negative mean absolute error.
- Random Forest Regression Pipeline: Data preprocessing is followed by a RandomForestRegressor, with GridSearchCV used to optimize key parameters such as number of estimators, tree depth, and minimum samples for splitting.

Table 2. GridSearchCV-Optimized Hyperparameters for Random Forest Regressor

| Hyperparameter   | Value          |
|------------------|----------------|
| n_estimator      | [50, 100, 200] |
| max_depth        | [None, 10, 20] |
| min_sample_split | [2, 5, 10]     |

In Table 2, the hyperparameters `n_estimators`, `max_depth`, and `min_samples_split` are presented. The `n_estimators` parameter specifies the number of decision trees constructed independently within the ensemble model, which are subsequently combined through a voting mechanism in classification tasks or an averaging process in regression tasks to generate the final prediction.

In this study, three primary hyperparameters were considered to optimize the performance of the ensemble learning model: `n_estimators`, `max_depth`, and `min_samples_split`. Each of these parameters influences the model's structure and learning behavior in distinct ways.

The `n_estimators` parameter determines the number of decision trees constructed independently within the ensemble. These trees are subsequently aggregated through a voting mechanism for classification tasks or an averaging process for regression tasks to produce the final prediction. A higher number of estimators generally enhances model stability and accuracy but also increases computational cost and training time.

The `max_depth` parameter defines the maximum allowable depth of each decision tree. This setting regulates the tree's complexity and its ability to capture intricate relationships in the dataset. In this study, three configurations None, 10, and 20 were evaluated. When set to None, the tree expands

without depth limitation, which can lead to overfitting. Restricting the depth to 10 produces a shallower structure that promotes generalization, whereas increasing it to 20 allows for more detailed learning at the expense of higher variance and computational demands.

Finally, the `min_samples_split` parameter specifies the minimum number of samples required to split an internal node. It governs how finely the model partitions the data. The values 2, 5, and 10 were tested to assess their effects on model performance. A smaller value such as 2 permits more frequent splits and results in deeper trees that may overfit, while larger values (e.g., 10) constrain growth, yielding simpler and more stable models that may risk underfitting.

Overall, fine-tuning these hyperparameters is essential to achieving a balance between model bias and variance, thereby improving the ensemble model's predictive capability and generalization performance

- In this pipeline, the SVR model was implemented with a parameter grid search to identify the optimal combination of the kernel type (linear or RBF), the regularization parameter (C), and the epsilon parameter that defines the margin of tolerance.

Table 3. SVR Hyperparameters Tuned Using GridSearchCV

| Hyperparameter | Value             |
|----------------|-------------------|
| Kernel         | ['linear', 'rbf'] |
| C              | [0.1, 1, 10]      |
| Epsilon        | [0.1, 0.2, 0.5]   |

In Table 3, the hyperparameters Kernel, C, and Epsilon are presented. The kernel parameter in Support Vector Regression (SVR) specifies the type of kernel function used to map input data into a higher-dimensional feature space, enabling the model to learn both linear and nonlinear relationships. In this study, two kernel types were utilized: linear and rbf (Radial Basis Function).

The C parameter in Support Vector Regression (SVR) regulates the balance between model complexity and the tolerance for errors in the training data. It serves as a regularization factor that controls how strictly the model fits the training samples. In this study, three values of C were examined: 0.1, 1, and 10.

A lower C value (0.1) applies stronger regularization, allowing a wider margin and greater tolerance for training errors. This configuration tends to produce a smoother and more generalized model but may lead to underfitting if the model becomes too simple. In contrast, a higher C value (10) reduces the regularization effect, forcing the model to minimize training errors and fit the data more closely, which may improve training accuracy but increase the risk of overfitting. The intermediate value  $C = 1$  typically provides a balanced trade-off between these two conditions.

Optimizing the C parameter is essential to achieving an appropriate balance between bias and variance, thereby enhancing the SVR model's ability to generalize effectively to unseen data.

The epsilon parameter in Support Vector Regression (SVR) defines the width of the epsilon-insensitive zone around the regression function, within which errors are not penalized. In other words, it specifies the margin of tolerance where deviations between the predicted and actual values are considered acceptable and do not contribute to the loss function.

In this study, three values of epsilon were examined: 0.1, 0.2, and 0.5. An epsilon value 0.1 creates a narrower margin, making the model more sensitive to small deviations and resulting in a closer fit to the training data, which can increase the risk of overfitting. Conversely, an epsilon value 0.5 establishes a wider margin of tolerance, allowing greater deviations without penalty and promoting a smoother regression function, though it may lead to underfitting if the margin is too broad. The value 0.2 provides a balanced compromise between sensitivity and generalization.

Proper tuning of the epsilon parameter is crucial for controlling the trade-off between model precision and robustness, ensuring that the SVR model captures meaningful trends while maintaining good generalization to unseen data.

- This pipeline employs a Multi-Layer Perceptron (MLP) model, which serves as an implementation of the Artificial Neural Network (ANN). A GridSearchCV procedure was applied to determine the optimal combination of hidden layer sizes, activation functions, and solvers for the model.

Table 4. Hyperparameter ANN Using GridSearchCV

| Hyperparameter    | Value              |
|-------------------|--------------------|
| hidden_layer_size | [(100,), (50, 50)] |
| activation        | ['relu', 'tanh']   |
| solver            | ['adam', 'sgd']    |

Table 4 presents the parameters and corresponding values used in the Multi-Layer Perceptron (MLP) model. The Multilayer Perceptron (MLP) model includes several key hyperparameters that define its architecture and optimization behavior, namely hidden\_layer\_sizes, activation, and solver. These parameters collectively influence the model's learning capacity, convergence speed, and ability to generalize.

The hidden\_layer\_sizes parameter specifies the number of neurons and layers in the network's hidden structure. In this study, two configurations were tested: (100,) and (50, 50). The first configuration, (100,), represents a single hidden layer containing 100 neurons, while (50, 50) denotes a deeper architecture with two hidden layers of 50 neurons each. Increasing the number of layers and neurons generally enhances the model's ability to learn complex relationships, though it may also increase the risk of overfitting and extend training time.

The activation parameter determines the nonlinear transformation applied to the output of each neuron. Two activation functions were considered: 'relu' (Rectified Linear Unit) and 'tanh' (hyperbolic tangent). The 'relu' function is widely used for its computational efficiency and ability to accelerate convergence by mitigating the vanishing gradient problem. In contrast, the 'tanh' function maps input values to a range between -1 and 1, which can be advantageous for normalized data but may lead to slower convergence.

The solver parameter defines the optimization algorithm used to adjust the model's weights during training. Two solvers were employed: 'adam' and 'sgd' (Stochastic Gradient Descent). The 'adam' solver combines the advantages of adaptive learning rate and momentum, making it effective for large or noisy datasets. Meanwhile, the 'sgd' solver updates weights incrementally per batch, providing greater control over learning dynamics but requiring careful tuning of the learning rate to ensure stable convergence.

Overall, tuning these hyperparameters network depth, activation function, and optimization algorithm is essential to achieving a balance between learning efficiency, model complexity, and generalization performance in MLP-based regression models.

### 3.3. Model Evaluation

At this stage, model evaluation was carried out using 5-fold cross-validation along with several performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination ( $R^2$ ), for each trained model—Linear Regression, Random Forest, Support Vector Regression (SVR), and Artificial Neural Network (ANN). These metrics were computed using the cross\_validate function, and their average values were reported for each model. The results provide an overview of model performance in terms of prediction error and the quality of fit to the data.

Cross-validation ensures a more objective evaluation by partitioning the dataset into multiple subsets and testing each subset in turn. Based on the cross-validation results obtained for the four models, a comparative analysis of their performance is presented below. Table 5 presents the results of the experiments.

Table 5. The Results of the Experiments.

| Model                     | MAE   | MSE   | RMSE  | R <sup>2</sup> |
|---------------------------|-------|-------|-------|----------------|
| Linear Regression         | 0.395 | 0.236 | 0.482 | 0.728          |
| Random Forest Regression  | 0.019 | 0.011 | 0.086 | 0.987          |
| Support Vector Regression | 0.370 | 0.260 | 0.500 | 0.731          |
| Artificial Neural Network | 0.253 | 0.107 | 0.324 | 0.868          |

The performance of four regression models—Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN)—was assessed using four statistical metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R<sup>2</sup>). These metrics provide a comprehensive view of each model's predictive accuracy and generalization ability.

As shown in Table 5, Random Forest Regression achieved the best results overall, with the lowest MAE (0.019), MSE (0.011), and RMSE (0.086), alongside the highest R<sup>2</sup> (0.987). This indicates that Random Forest can make highly accurate predictions and effectively capture complex nonlinear relationships in the data. The ANN also demonstrated strong performance, with relatively low error values (MAE = 0.253, RMSE = 0.324) and a high R<sup>2</sup> (0.868). While not as precise as Random Forest, the ANN was able to learn meaningful patterns and generalize well to new data. In contrast, Linear Regression and SVR showed higher errors and lower R<sup>2</sup> values (0.728 and 0.731, respectively), suggesting these models struggle to capture the nonlinear relationships present in the dataset, resulting in less accurate predictions.

Overall, the evaluation indicates that Random Forest Regression is the most reliable approach for this dataset. The ANN offers a solid alternative with strong learning capability, whereas Linear Regression and SVR provide moderate performance.

#### 4. DISCUSSIONS

The experimental results provide a comprehensive comparison of the predictive capabilities of four regression models: Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN). As summarized in Table 5, the evaluation revealed significant differences in accuracy and generalization performance among the models.

Among the tested models, Random Forest Regression demonstrated the highest predictive performance, achieving the lowest errors (MAE = 0.019; MSE = 0.011; RMSE = 0.086) and the highest coefficient of determination (R<sup>2</sup> = 0.987). This suggests that the ensemble learning approach of Random Forest, which combines multiple decision trees through averaging, effectively reduces variance and enhances model robustness. Its ability to capture nonlinear relationships and complex interactions among features makes it particularly suitable for this dataset.

The Artificial Neural Network (ANN) also performed well, showing moderate error values (MAE = 0.253; RMSE = 0.324) and a high R<sup>2</sup> (0.868). The multilayer structure of ANN allows it to model nonlinear relationships in the data, providing better generalization than linear models. However, its accuracy is slightly lower than Random Forest, likely due to sensitivity to hyperparameter settings, training iterations, or potential overfitting in certain configurations.

In contrast, Linear Regression and SVR produced higher errors and lower R<sup>2</sup> scores (0.728 and 0.731, respectively). Linear Regression, constrained by its linearity assumption, struggled to model the

**Commented [A7]:** Add comparison with previous study.  
Add more explanation why RFR works better for this dataset.  
Add practical implications for Hospitals.  
Discuss the limitations.

complex relationships in the dataset. SVR performance was likely affected by the choice of kernel and regularization settings, which are critical for capturing nonlinear patterns effectively.

The model exhibits considerable potential for further refinement; nevertheless, it requires evaluation by domain experts or end-users, as well as testing on new real-world scenarios to ensure its robustness and applicability.

## 5. CONCLUSION

Based on the comparison of Linear Regression, Random Forest Regression, Support Vector Regression, and Artificial Neural Network algorithms for developing a model to predict the length of stay (LOS) of newborns, the Random Forest Regression method demonstrated the highest performance. The study results indicate that both medical and non-medical factors—such as initial medical condition, primary diagnosis, age, gender, and referral type affect the length of hospitalization for newborn patients. Random Forest Regression showed the best performance, even when applied to a relatively small dataset.

A limitation of this study is the small size of the dataset, which led to suboptimal performance for some of the methods used. Therefore, future research is recommended to incorporate a larger dataset and to explore additional variables, including other medical factors such as birth weight, gestational age, and parental history, as well as non-medical factors such as hospital facilities and geographic location.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the research, authorship, or publication of this article. All stages of the study, including data analysis and manuscript preparation, were conducted objectively without any financial, professional, or institutional influence.

## REFERENCES

- [1] WHO, "Newborn Mortality," WHO. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021>
- [2] Y. G. Robi and T. M. Sitote, "Neonatal Disease Prediction Using Machine Learning Techniques," *J. Healthc. Eng.*, vol. 2023, no. 1, pp. 1–16, 2023, doi: 10.1155/2023/3567194.
- [3] G. Almeida, F. B. Correia, and A. R. Borges, "Hospital Length-of-Stay Prediction Using Machine Learning Algorithms — A Literature Review," *Applied Sci.*, vol. 2024, no. 14, pp. 1–20, 2024, doi: <https://www.mdpi.com/2076-3417/14/22/10523>.
- [4] L. Garg *et al.*, "Characterising Hospital Admission Patterns and Length of Stay in the Emergency Department at Mater Dei Hospital Malta," *Preprints*, vol. 1, no. February, pp. 28–30, 2023, doi: 10.20944/preprints202302.0315.v1.
- [5] K. Stone, R. Zwigelaar, P. Jones, N. Mac Parthaláin, and D. Yoon, "A Systematic Review of The Prediction of Hospital Length of Stay: Towards a Unified Framework," *PLOS Digit. Heal.*, vol. 1, no. 4, pp. 1–39, 2022, doi: 10.1371/journal.pdig.0000017.
- [6] M. A. Shbool *et al.*, "Machine Learning Approaches to Predict Patient's Length of Stay in Emergency Department," *Appl. Comput. Intell. Soft Comput.*, vol. 2023, no. 1, pp. 1–13, 2023, doi: 10.1155/2023/8063846.
- [7] S. G. Gurazada, S. Gao, F. Burstein, and P. Buntine, "Predicting Patient Length of Stay in Australian Emergency Departments Using Data Mining," *Sensors*, vol. 22, no. 13, pp. 1–15, 2022, doi: <https://www.mdpi.com/1424-8220/22/13/4968>.
- [8] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, and A. Duclos, "The Prediction of Hospital Length of Stay using Unstructured Data," *BMC Med. Inform. Decis. Mak.*, vol. 4, pp. 1–9, 2021, doi: 10.1186/s12911-021-01722-4.
- [9] S. Barnes, E. Hamrock, M. Toerper, S. Siddiqui, and S. Levin, "Real-Time Prediction of Inpatient Length of Stay for Discharge Prioritization," *Res. Appl.*, vol. 23, pp. 2–10, 2016, doi: 10.1093/jamia/ocv106.

Commented [A8]: Add explicit contribution statement.

Commented [A9]: Some DOIs missing and the journal names not in full format. Fix it.  
WHO website citation should not dominate.  
Correct spelling errors in the references.  
Preprints references should be minimized.

- [10] X. Zeng, "Length of Stay Prediction Model of Indoor Patients Based on Light Gradient Boosting Machine," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–14, 2022.
- [11] L. Chen and H. B. Klasky, "Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients with Sepsis : A Comparative Study," *SoutheastCon 2022*, vol. 2022, pp. 302–309, 2022.
- [12] J. Wu *et al.*, "Development of A Scoring Tool for Predicting Prolonged Length of Hospital Stay in Peritoneal Dialysis Patients Through Data Mining," *Ann. Transl. Med.*, vol. 8, no. 21, pp. 1–15, 2020, doi: 10.21037/atm-20-1006.
- [13] J. An, M. Jung, S. Ryu, Y. Choi, and J. Kim, "Analysis of Length of Stay for Patients Admitted to Korean Hospitals Based on The Korean National Health Insurance Service Database," *Informatics Med. Unlocked*, vol. 37, no. October 2022, p. 101178, 2023, doi: 10.1016/j.imu.2023.101178.
- [14] A. I. Taloba, R. M. A. El-aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning," *J. Healthc. Eng.*, vol. 2022, no. 1, pp. 1–10, 2022.
- [15] S. Melhem, A. Al-Aiad, and M. S. Al-Ayyad, "Patient Care Classification Using Machine Learning Techniques," in *International Conference on Information and Communication Systems (ICICS)*, 2021, pp. 1–6. doi: 10.1109/ICICS52457.2021.9464582.
- [16] F. Chollet, *Deep Learning with Python*. United States: Manning Publications Co, 2018. doi: 10.4018/978-1-7998-3095-5.ch003.
- [17] A. Masbakhah, U. Sa'adah, and M. Muslikh, "Heart Disease Classification Using Random Forest and Fox Algorithm as Hyperparameter Tuning," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 4, pp. 964–976, 2025, doi: 10.35882/jeeemi.v7i4.932.
- [18] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, no. April 2015. 2015. doi: 10.1007/978-1-4302-5990-9.
- [19] H. Huang, R. Jia, J. Liang, J. Dang, and Z. Wang, "Wind Power Deterministic Prediction and Uncertainty Quantification Based on Interval Estimation," *J. Sol. Energy Eng. Trans. ASME*, vol. 143, no. 6, pp. 1–12, 2021, doi: 10.1115/1.4051430.
- [20] L. Costa *et al.*, "Multilayer Perceptron," in *Introduction to Computational Intelligence*, no. January, IEEE, 2023, pp. 1–6. doi: 10.1145/3090354.3090427.
- [21] H. Yin, "Enhancing Directional Accuracy in Stock Closing Price Value Prediction Using a Direction-Integrated MSE Loss Function," in *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, 2023, pp. 119–126. doi: 10.5220/0012810200003885.
- [22] T. O. Hodson, "Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Tot," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
- [23] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting Health Care Costs Using Evidence Regression," in *MDPI*, 2019, pp. 1–13. doi: 10.3390/proceedings2019031074.
- [24] D. Sivasundaralingam and N. Thasavaran, "Predicting Inpatient Bed Demand Using Machine Learning," in *YOUNG MEMBERS' TECHNICAL CONFERENCE 2024*, 2023, pp. 150–157. [Online]. Available: [https://www.researchgate.net/publication/391584272\\_Predicting\\_Inpatient\\_Bed\\_Demand\\_Using\\_Machine\\_Learning](https://www.researchgate.net/publication/391584272_Predicting_Inpatient_Bed_Demand_Using_Machine_Learning)
- [25] S. Al-Gahtani and M. M. Shoukri, "Analysis of Length of Stay ( LOS) Data from the Medical Records of Tertiary Care Hospital in Saudi Arabia for Five Diagnosis Related Groups : Application of Cox Prediction Model," *Open J. Stat.*, vol. 11, pp. 99–112, 2021, doi: 10.4236/ojs.2021.111005.
- [26] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-nejad, "Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation," *JMIR Med. INFORMATICS*, vol. 9, no. 5, pp. 1–23, 2021, doi: 10.2196/21347.
- [27] D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon, "Predicting Inpatient Flow at A Major Hospital using Interpretable Analytics," *Manuf. Serv. Oper. Manag.*, vol. 2020, no. 1, pp. 1–42,

2020, doi: <https://doi.org/10.1101/2020.05.12.20098848>.

#### SUBMISSION CHECK-LIST (MUST BE INCLUDED IN ARTICLE FILE)

All items must be YES before submitting to Jurnal Teknik Informatika (JUTIF) in order to make sure a manuscript well-prepared within a high standard. Please put X in column Yes after checking each item

| General Format          |                                                                                                                                                           |     |    |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|----|
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | I have prepared the manuscript using "JUTIF-Template.docx"                                                                                                | X   |    |
| 2                       | The manuscript is written in single or double column using Times New Roman 11 pt font                                                                     | X   |    |
| 3                       | Title is matching with the focus of the manuscript                                                                                                        | X   |    |
| 4                       | I have complete authors information:                                                                                                                      | X   |    |
|                         | a. Affiliation of all authors                                                                                                                             | X   |    |
|                         | b. Indicate one of the authors is the corresponding author                                                                                                | X   |    |
|                         | c. Email of corresponding author                                                                                                                          | X   |    |
| 5                       | The length of the manuscript is not less than 3500 words (excluding abstract, author information and reference list)                                      | X   |    |
| 6                       | The manuscript contains the core contents: introduction with literature review, method, results, discussion, conclusion, conflict of interest, references | X   |    |
| 7                       | Contains maximum 2 level of sections                                                                                                                      | X   |    |
|                         |                                                                                                                                                           |     |    |
| Abstract                |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | The abstract is written between 150 – 250 words                                                                                                           | X   |    |
| 2                       | contains the purpose                                                                                                                                      | X   |    |
| 3                       | contains the method / research approach                                                                                                                   | X   |    |
| 4                       | contains the results / conclusion                                                                                                                         | X   |    |
| 5                       | contains novelty/originality/important findings/impact                                                                                                    | X   |    |
| 6                       | The abstract is completed with keywords                                                                                                                   | X   |    |
|                         |                                                                                                                                                           | X   |    |
| Figures (if applicable) |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | All figures are clear and readable (image, text and legend) with high resolution                                                                          | X   |    |
| 2                       | All figures are in formal style, without redundant title                                                                                                  | X   |    |
| 3                       | All information in the figure are in English and all decimal written in international standard, using point (.) not comma (,)                             | X   |    |
| 4                       | All figures have captions (at the bottom) with consecutive numbers                                                                                        | X   |    |
| 5                       | All figures are cited in the text using consistent citation style                                                                                         | X   |    |
| 6                       | Never citing figure using below and above, but using the number of figure                                                                                 | X   |    |
| Tables (if applicable)  |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | All texts in tables are clear and readable                                                                                                                | X   |    |
| 2                       | All tables are in formal style                                                                                                                            | X   |    |

|                                  |                                                                                                                              |            |           |
|----------------------------------|------------------------------------------------------------------------------------------------------------------------------|------------|-----------|
| 3                                | All information in the table are in English and all decimal written in international standard, using point (.) not comma (,) | X          |           |
| 4                                | All tables are captioned on top with consecutive numbers                                                                     | X          |           |
| 5                                | All tables are cited in the text using consistent citation                                                                   | X          |           |
| 6                                | Never citing table using below and above, but using the number of table                                                      | X          |           |
| <b>Equations (if applicable)</b> |                                                                                                                              |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                  | <b>Yes</b> | <b>No</b> |
| 1                                | All equations are written using editor tool (editable), not a cropped image                                                  | X          |           |
| 2                                | All equations are captioned on top with consecutive numbers                                                                  | X          |           |
| 3                                | All equations are cited in the text using consistent citation                                                                | X          |           |
| 4                                | Never citing equation using below and above, but using the number of equation                                                | X          |           |
| <b>References</b>                |                                                                                                                              |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                  | <b>Yes</b> | <b>No</b> |
| 1                                | Reference list and citation consistently follows the IEEE style                                                              | X          |           |
| 2                                | Journal names in the reference list are not in abbreviated format, but it should be in full name format                      | X          |           |
| 3                                | All references are cited in the text                                                                                         | X          |           |
| 4                                | Citation in the text follows general consistent citation rules                                                               | X          |           |
| 5                                | Paper cites at least 25 references                                                                                           | X          |           |
| 6                                | Book sources are not more than 20% of the reference list                                                                     | X          |           |
| 7                                | 80% of references should come from journals/conferences                                                                      | X          |           |
| 8                                | If references come from journals, it should mention name of journal, volume, no, page, and DOI.                              | X          |           |
| 9                                | Self-citation in the list is not more than 3                                                                                 | X          |           |
| 10                               | Write all authors in the reference list unless authors more than 7 (Write the first 6, then et al.)                          | X          |           |

## 7. Hasil Review dari Reviewer 2 pada File

Jurnal Teknik Informatika (JUTIF)

P-ISSN: 2723-3863

E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 1530-1537

<https://jutif.if.unsoed.ac.id>

DOI: <https://doi.org/10.52436/1.jutif.6.6.3540>

### Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025

Phone Number : xxxxxxx (Whatsapp/Cellphone Number, For Corresponding)

The cellphone number is only for ease of communication and will NOT be displayed in the article

#### Abstract

The length of stay (LOS) of newborns is a crucial indicator for healthcare management and hospital resource allocation. However, the factors influencing LOS are not yet fully understood, and existing prediction methods often yield suboptimal results. This study aims to identify the factors affecting LOS and to develop a predictive model for newborn LOS using several machine learning algorithms. A comparison was conducted among Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN) to determine the most effective model. The data were obtained from the medical records of newborn patients at three private hospitals in Indonesia. The research followed several stages, including data collection and understanding, data preprocessing, modeling, and evaluation. The experimental results demonstrated that the Random Forest Regression algorithm achieved the best predictive performance, with MAE = 0.019, MSE = 0.011, RMSE = 0.086, and  $R^2 = 0.987$ . The feature importance analysis of the Random Forest model revealed that gender, referral source, insurance type, and diagnosis were the most influential factors affecting LOS. Theoretically, this study strengthens the application of machine learning in healthcare data analytics, particularly in predicting the length of stay for newborn patients.

**Keywords :** length of stay, machine learning, newborns, prediction, Random Forest;

This work is an open access article licensed under a Creative Commons Attribution 4.0 International License.



## 1. INTRODUCTION

Newborn health serves as a key indicator for assessing the quality of a country's healthcare services. Infant mortality and morbidity rates reflect not only the health status of mothers and children but also the overall effectiveness of the healthcare system [1]. As reported by Statistics Indonesia in 2020, although Indonesia has experienced a decline in infant mortality over the past few decades, regional disparities remain significant, indicating inequality in access to and quality of healthcare services. Certain regions, such as Maluku and Papua, continue to record higher infant mortality rates compared to the national average, while areas like Java are approaching the *Sustainable Development Goals* (SDGs) target of 12 deaths per 1,000 live births.

Accurately predicting the LOS of newborns holds significant implications for hospitals in terms of resource management and cost efficiency [2], [3]. Reliable prediction models enable healthcare providers to design more effective care strategies, optimize hospital bed utilization, and reduce patients' financial burdens [3], [4]. However, traditional approaches such as logistic regression and survival analysis often fail to adequately address the non-linear and complex nature of medical data [5], [6].

One of the critical aspects of neonatal care is the length of hospital stay (LOS), which varies considerably depending on several medical factors such as prematurity, respiratory disorders, infections, and congenital anomalies [7]. Prematurity, for example, is a major contributor to extended hospitalization due to the incomplete development of vital organs, particularly the lungs and immune system [3]. Similarly, Respiratory Distress Syndrome (RDS) is frequently cited as a primary reason for prolonged admission in neonatal intensive care units [8]. In addition, LOS is also influenced by age, gender, and social determinants [4].

**Commented [A1]:** Judul sudah sesuai, yaitu sudah mengandung objek penelitian, metode dan tujuan penelitian.  
Tidak menggunakan singkatan atau istilah umum yang tidak dijelaskan.

**Commented [A2]:** Abstrak sudah memuat masalah pada penelitian, tujuan, metode dan hasil penelitian.  
Saran penulisan dampak atau urgensi penelitian agar lebih eksplisit.  
Kata kunci sudah berisi 5 frase yang berkaitan dengan isi penelitian dan disusun secara alphabet.

With the advancement of data-driven technologies, machine learning (ML) has emerged as a promising alternative for predictive analytics in healthcare. Numerous studies have demonstrated its potential in forecasting hospital stay durations. Almeida [3] reported that the Random Forest algorithm can produce highly accurate predictions when optimal feature selection is applied. Barnes [9] utilized real-time data from pediatric intensive care units. Moreover, Zeng [10] found that LightGBM achieved superior predictive performance, and Shbool et al. [6], [11] demonstrated that Random Forest can effectively enhance hospital resource utilization, particularly during the COVID-19 pandemic.

In this study, several machine learning algorithms—Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—are employed to predict neonatal LOS. These algorithms were selected for their respective strengths in handling data with varying levels of complexity. LR serves as the baseline model to capture linear relationships between input and output variables. RFR is chosen for its robustness in modeling non-linear interactions and combining multiple decision trees to enhance predictive accuracy [6], [11], [12]. SVR offers flexibility in capturing non-linear patterns through the use of kernel functions, while ANN is utilized for its capability to learn complex and high-dimensional representations in medical datasets, thereby improving predictive precision [13].

Model performance is evaluated using four quantitative metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared ( $R^2$ ). MAE measures the average magnitude of errors in the same units as the data, MSE penalizes larger deviations more heavily, RMSE retains interpretability while accounting for outliers, and  $R^2$  assesses the proportion of variance explained by the model. Collectively, these metrics provide a comprehensive framework for assessing model performance [10], [11].

The primary objective of this study is to identify key determinants influencing the LOS of newborns and to develop more accurate predictive models compared to conventional statistical methods. Understanding these determinants is crucial for hospitals to optimize resource allocation—including bed capacity, medical personnel, and neonatal care equipment. The analysis incorporates both medical and non-medical factors, such as gender, clinical conditions (e.g., prematurity, respiratory distress, or infection), and socio-economic characteristics that may affect hospitalization outcomes.

Furthermore, this study seeks to compare the predictive performance of different machine learning models in capturing the non-linear structures inherent in medical data. The evaluation using MAE, MSE, RMSE, and  $R^2$  provides a detailed comparison of the predictive reliability and explanatory power of each algorithm. Overall, this research not only advances predictive modeling in healthcare but also supports the broader goal of improving the quality of care for newborn patients in hospital settings.

## 2. METHOD

To ensure consistency throughout the research process, this study was structured into several stages that align with its objectives. The main phases of the research include data collection, data preprocessing, modeling, and model evaluation. Each of these stages is described in detail in the following subsections.

### 2.1. Data Collection

The first step of this research involves understanding the business requirements and primary objectives of the study. Hospitals often face challenges in managing bed capacity and medical resources, particularly for newborn patients who require specialized care. Therefore, this study aims to develop a predictive model to estimate the length of stay (LOS) based on both medical and non-medical factors. By utilizing such a model, hospitals can plan resource allocation more effectively and enhance service efficiency.

**Commented [A3]:** Pendahuluan sudah menyajikan masalah atau pentingnya penelitian. Mohon tambahkan gap dari penelitian yang dibandingkan dengan penelitian lain secara jelas. Rujukan pada pendahuluan minimal 25, maks 5 tahun terakhir. Perlu penambahan penegasan novelty di akhir sebelum tujuan. Tujuan belum ditulis eksplisit di akhir paragraf.

The study uses data from newborn patients hospitalized in selected hospitals, encompassing clinical and demographic variables such as infant initial diagnosis, medical interventions, and other relevant medical and non-medical factors influencing the duration of hospitalization.

1. Population: All newborns admitted to three hospitals during the study period.
2. Inclusion Criteria: Newborns with complete medical records who were hospitalized during the year 2023.
3. Exclusion Criteria: Newborns with incomplete medical records or those referred to other hospitals before completing treatment.

## 2.2. Data Preprocessing

This stage involves several data preprocessing steps to ensure the dataset is in optimal condition before being used for machine learning modeling:

- Data Deduplication: Removing duplicate entries to maintain the accuracy and integrity of the dataset.
- Data Cleansing: Handling missing values using imputation methods (mean or median) and removing inconsistent or extreme data points.
- Data Transformation: Applying normalization to numerical features and encoding categorical features to ensure the data is properly formatted for model training.
- Dimensionality Reduction: Reducing the dimensionality of diagnostic features by utilizing ICD-10 codes, extracting the corresponding *Block* and/or *Chapter* levels for feature representation.
- Data Splitting: Dividing the dataset into a training set (80%) and a testing set (20%) to ensure fair model training and evaluation. Additionally, cross-validation techniques are employed to improve model generalization.

## 2.3. Modeling

Based on the studies that have been conducted, this study uses four machine learning algorithms to build a prediction model for the length of hospitalization, namely LR, RFR, SVR and ANN.

### a. Linear Regression

As a supervised machine learning technique, linear regression is trained on labeled datasets to construct the best-fit linear function. This optimized function aims to capture the linear relationship the idea that the output shifts uniformly (at a constant rate) whenever the input changes and is subsequently employed for forecasting values on unseen data. The relationship itself is fundamentally depicted by a straight line [14], [15]. The model of linear regression can be expressed in formula (1).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

where Y represents the dependent variable, which corresponds to the predicted output, while X denotes the independent variable or predictor.  $\beta_0$  refers to the intercept (constant term),  $\beta_1$  is the regression coefficient associated with X, and  $\varepsilon$  represents the error term.

In simple linear regression, the regression coefficients  $\beta_0$  and  $\beta_1$  can be calculated using the Ordinary Least Squares (OLS) method with the formula:

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad (3)$$

where  $X_i$  and  $Y_i$  represent the dependent and independent variables, respectively.  $\bar{X}$  denotes the mean value of X,  $\bar{Y}$  denotes the mean value of Y, and n refers to the total number of data points.

**Commented [A4]:** Persamaan diberikan penomoran dan dirujuk pada teks.

If there is more than one independent variable, then the model becomes multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (4)$$

where  $Y$  is the dependent variable, namely the predicted output.  $\beta_0$  is an intercept or constant.  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficient for each independent variable.  $X_1, X_2, \dots, X_n$  are an independent variable, namely the predictor.  $\varepsilon$  is an error.

The following are the steps in the modeling process using a linear regression algorithm in a study on predicting the length of hospitalization for newborns:

- Determine the input variables ( $X_1 \dots X_n$ ), such as gender, initial diagnosis, and insurance, which are suspected of influencing  $Y$  (length of hospitalization).
- Standardize or rescale the data to avoid bias due to differences in units between features before the regression process.
- Apply a linear model to map the relationship between  $X_1, \dots, X_n$  and  $Y$  using the OLS approach.
- Calculate the values of  $\beta_0, \beta_1, \dots, \beta_n$  that minimize the residual error  $\varepsilon$ , so that the regression line approximates the actual patient length of stay.

#### b. Random Forest Regression

Random Forest Regression (RFR) is a powerful ensemble machine learning technique that utilizes a collection of decision trees to achieve a more robust and accurate prediction. The central goal of RFR is to prevent the high variance and overfitting issues typical of single decision trees [16].

It achieves this by injecting randomness during the training process:

- Each tree is built using a bootstrap sample (a random subset with replacement) of the original data.
- At every node split, only a random subset of features is considered.

By combining and averaging the predictions from these diverse, independently-built trees, the final RFR model improves generalization and stability. Furthermore, Out-of-Bag (OOB) samples—the data excluded from each tree's training set—are used for internal performance validation.

The use of RFR learning ensembles, capable of producing models with low variability and higher accuracy. Eq. (3) represent RFR:

$$R(a) = \sum_{k=1}^N Y_k \sum_{j \in C_k} \theta(j; p_k) = \sum_{k=1}^N \sum_{j \in C_k} Y_k \theta(j; p_k) \quad (5)$$

with  $R(a)$  is the final decision of the classifier,  $Y_k$  average response value ke-k,  $C_k, P_k$  is a divisive parameter to divide the region of the decision and  $\theta(j; p_k)$  is a function of the information limit based on the  $P_k$  and  $j$ -index. This equation shows the aggregation of the response values in each region  $k$ , taking into account the data subes and relevant divisor parameters [17].

#### c. Support Vector Regression

Support Vector Regression (SVR) is an extension of SVM designed to address regression problems. In SVM, the primary goal is to find the best hyperplane that separates data classes. Conversely, in SVR, the primary goal is to find a function that predicts continuous values based on the input, while maintaining the SVM's margin principles. SVR works by building a model that not only accounts for prediction errors but also minimizes model complexity. To do this, SVR uses an epsilon-insensitive loss function, which allows small errors to be ignored if they are within an epsilon distance from the predicted value [18].

The Support Vector Regression (SVR) modeling process consists of several key steps. First, a mathematical model is constructed to map input variables, such as diagnosis and gender, to the target variable, namely the length of stay. A tolerance margin is then established to define the acceptable range of prediction errors that do not incur penalties. Errors falling within this range are disregarded to reduce overfitting, while those exceeding the tolerance threshold receive larger penalties. The model is subsequently optimized to maintain a balance between simplicity and predictive accuracy. In cases where the relationships between variables are non-linear, kernel functions are applied to transform the data into a higher-dimensional space, thereby facilitating better pattern recognition. Support vectors—data points located on or beyond the tolerance boundaries—are then identified, as they play a crucial role in determining the structure of the regression function. Finally, the trained SVR model is used to predict the length of stay for new patient data.

SVR maps a low-dimensional input space into a high-dimensional feature space using a nonlinear mapping and then performs a linear regression in the high-dimensional feature space. Given the input dataset  $D = \{(x_i, y_i) : i = 1 \cdots N\}$ , the wind power prediction of SVR method can be given as  $y = f(x) = \omega \phi(x) + b$ , where  $w$  is the weight,  $b$  is the parameter of bias, and  $\phi(x)$  represents the high-dimensional space. To obtain parameters  $w$  and  $b$ , the prediction of SVR model with structure risk minimization principle can be transformed to a mathematical optimization problem and solved using Lagrange multiplier [19]. The optimization problem is presented as follows:

$$f(X, Y_i, Y_i^*) = \sum_{i=1}^N (Y_i - Y_i^*) K(X, X_i) + b \quad (6)$$

where  $Y_i$  and  $Y_i^*$  are the Lagrange multiplier and  $K(X, X_i)$  represents the kernel function. In this study, the radial-basis function is adopted as the kernel function of SVR. Some important parameters of SVR-like kernel coefficient, penalty parameter of the error term and radius are selected by many trials with empirical data.

#### d. Artificial Neural Network

An Artificial Neural Network (ANN) is a computer model inspired by how the human brain processes information. It is made up of layers of connected neurons, which allow the network to learn more complex patterns in data as it trains. ANNs work by transforming input data through multiple layers, with each layer learning to extract more meaningful features. The network is trained using backpropagation, a process that adjusts the connection weights based on the difference between the network's output and the expected target [20].

A common ANN structure is the multi-layer backpropagation network, which has three types of layers: the input layer, hidden layers, and output layer. The input layer receives the input variables ( $X_1, X_2, \dots, X_n$ ). The hidden layers process this information ( $Z_1, Z_2, \dots, Z_n$ ), while the output layer produces the final results ( $Y$ ), such as sample classifications.

The network updates the weights of each layer to reduce errors. The error shows how well the network is learning—it is calculated as the difference between the network's current output and the desired output.

During forward propagation, the network calculates the weighted sum of inputs for each neuron. If the input signal passes a certain threshold, it goes through an activation function, which determines the neuron's output. Activation functions must be continuous, differentiable, and non-decreasing to work with backpropagation. A common example is the Binary Sigmoid function (as shown in Equations (7) and (8)).

$$Y = f(X) = \frac{1}{1 + e^{-\sigma x}} \quad (7)$$

$$f(X) = \sigma f(X)[1 - f(X)] \quad (8)$$

The following explains how the multilayer perceptron model operates using backpropagation.

Step 0: Initialize the weights, preferably set to small random values,

Step 1: If the condition is not met, do steps 2-9,

Step 2: For each training pair, perform steps 3-8.

Steps 3-5 are called forward propagation.

Step 3: Each input unit ( $X_i, i = 1, 2, \dots, n$ ) receives the  $X_i$  signal and transmits this signal to all units in the layer above it (hidden units),

Step 4: Each hidden unit ( $X_i, i = 1, 2, \dots, p$ ) sums the weights of its input signals,

$$Z\_in_j = V_{oj} + \sum_{i=1}^n X_i V_{ij} \quad (9)$$

$V_{oj}$  = bias on hidden unit  $j$  apply its activation function to calculate its output signal,

$Z_j = f(Z\_in_j)$ , and send this signal to all units in the layer above it (output units).

Step 5: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ), summing the weights of its input signals,

$$Y\_in_k = W_{ok} + \sum_{j=1}^n Z_j W_{jk} \quad (10)$$

$W_{ok}$  = bias on the output unit  $k$  and apply its activation function to calculate its output signal,  $Y_k = f(Y\_in_k)$ .

In Steps 6-8 it is called backpropagation:

Step 6: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ) receive the interconnected target patterns at the input of the training patterns, calculate the information error,

$$\Delta_k = (T_k - Y_k) f^1(T\_in_k) \quad (11)$$

calculate the weight correction (used to update the  $W_{jk}$  later),

$$\Delta W_{jk} = \alpha \Delta_k Z_j \quad (12)$$

calculate the bias correction (used to update the  $W_{ok}$  later), and send  $\delta_k$  to the units in the layer below,

Step 7: Each layer unit is hidden ( $Z_j, j = 1, 2, \dots, p$ ) sum the results of the input changes (from the units of the layers above),

$$\Delta\_in_j = \sum_{k=1}^n \Delta_k W_{jk} \quad (13)$$

multiply by the derivative of its activation function to calculate its error information,

$$\Delta_j = \Delta\_in_j f^1(Z\_in_j) \quad (14)$$

calculate the weight correction (used to update  $V_{oj}$  later),

Step 8: Each output unit ( $Y_k, k = 1, 2, \dots, m$ ) update bias and weights ( $j = 0, 1, \dots, p$ ):

$$W_{jk}(\text{new}) = W_{jk}(\text{old}) + \Delta W_{jk} \quad (15)$$

Each hidden layer unit ( $Z_j, j = 1, 2, \dots, p$ ) update bias and weights ( $i = 0, 1, 2, \dots, n$ ):

$$V_{ij}(\text{new}) = V_{ij}(\text{old}) + \Delta V_{ij} \quad (16)$$

Step 9: Test the stopping condition.

Steps of the testing process:

Following the training phase, the backpropagation neural network is applied strictly using its forward propagation stage. The application steps are as follows:

Step 0: Initialize weights (from the training algorithm).

Step 1: for each input vector, perform the steps 2-4.

Step 2: for  $i = 1, 2, 3, \dots, n$ : set the activation of the  $X_i$  input unit.

Step 3: for  $j = 1, 2, \dots, p$ :

$$Z\_in_j = V_{oj} + \sum_{i=1}^n X_i V_{ij} \quad (17)$$

$$Z_j = f(Z\_in_j) \quad (18)$$

Step 4: for  $k = 1, 2, \dots, m$ :

$$Y\_in_k = V_{ok} + \sum_{j=1}^n X_j V_{jk} \quad (19)$$

$$Y_k = f(Z\_in_k) \quad (20)$$

Step 5: if  $Y_k \geq 0,5$  then  $Y_k = 1$ , else  $Y_k = 0$ .

## 2.4 Evaluation Model

After modeling, the next step is to test the model's performance using various evaluation metrics, namely MSE, RMSE, MAE, and  $R^2$ . This testing is done to determine which model provides the most accurate predictions.

Mean Squared Error (MSE) is an equation that measures the average squared error between actual and predicted values. A smaller MSE value indicates better model performance (predictions are closer to reality) [14], [21], as expressed in the following equation:

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (21)$$

where  $n$  merupakan sample datam  $y_i$  represents the true value,  $\hat{y}_i$  denotes the predicted value, and  $(y_i - \hat{y}_i)^2$  denotes the squared difference between the actual and predicted values.

The Root Mean Squared Error (RMSE) is a widely used statistical metric that measures the average magnitude of the prediction errors in a regression model. It is defined as the square root of the mean of the squared differences between the actual and predicted values [14], [22], as expressed in the following equation:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (22)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Mean Absolute Error (MAE) is a commonly used evaluation metric that measures the average magnitude of the errors in a set of predictions, without considering their direction. It is defined as the mean of the absolute differences between the actual and predicted values [14], [22], as expressed in the following equation:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (23)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Coefficient of Determination ( $R^2$ ), commonly referred to as R-squared, is a statistical metric used to evaluate the goodness of fit of a regression model. It represents the proportion of the variance in the actual (observed) data that is explained by the predicted values generated by the model [23], [24]. The mathematical formulation of  $R^2$  is expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (24)$$

where  $y_i$  denotes the actual value,  $\hat{y}_i$  represents the predicted value,  $\bar{y}$  is the mean of the actual values, and  $n$  is the total number of observations. The numerator  $\sum (y_i - \hat{y}_i)^2$  represents the Residual Sum of Squares (RSS), while the denominator  $\sum (y_i - \bar{y})^2$  represents the Total Sum of Squares (TSS).

An  $R^2$  value close to 1 indicates that the model explains most of the variability in the data, signifying a strong predictive capability. Conversely, an  $R^2$  value near 0 suggests that the model fails to capture the underlying data pattern. In some cases,  $R^2$  can be negative if the model performs worse than a simple mean-based prediction.

**Commented [A5]:** Metode belum menampilkan tahapan penelitian menggunakan gambar.  
Setiap gambar dan tabel diberikan penjelasan dan disebut penomorannya dengan benar pada teks.  
Penulisan equation mohon cantumkan angka penomorannya di samping kanan rumus, dan dirujuk pada teks. Sesuaikan dengan template Jutif.  
Mohon penjelasan proses feature engineering (ICD-10 grouping) dapat diperjelas.  
Perlu penjelasan tipe data (numeric/categorical) + total fitur model.

### 3. RESULT

The study results are organized into three subsections: Data Collection and Preprocessing, Modeling, and Model Evaluation. The first subsection details the dataset and preprocessing outcomes, the second compares the RL, RFR, SVR, and ANN algorithms, and the third presents the evaluation results.

#### 3.1. Data Collection and Preprocessing

The length of stay (LOS) for newborn infants plays a vital role in planning medical care and managing hospital resource allocation. This study aims to determine the factors influencing neonatal LOS and to develop a machine learning-based prediction model that enhances efficiency and accuracy in hospital management. The data used in this research were collected from three private hospitals in Indonesia.

The dataset successfully collected comprised 3,244 cases obtained from medical records. The data included inpatient information with attributes such as gender, type of insurance, diagnosis, referral procedure, referral type, and length of stay (LOS). Following data cleaning and validation in collaboration with hospital physicians, a refined dataset containing 445 cases was produced. This cleaned dataset was subsequently utilized for model development. Table 1 provides an illustration of the dataset employed in predicting the length of stay (LOS) of newborn patients.

Table 1. Example of a Dataset Used for Predicting the Length of Stay (LOS) of Newborns

| No  | Gender | Diagnosis | Insurance Claim | Referral Procedure                          | Referral Patient | LOS |
|-----|--------|-----------|-----------------|---------------------------------------------|------------------|-----|
| 1   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 2   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 3   | Male   | P22.9     | Government      | Umbilical venous catheter                   | Yes              | 2   |
| 4   | Female | P07.2     | Private         | Parenteral nutrition                        | Yes              | 1   |
| 5   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 5   |
| ... | ...    | ...       | ....            | ....                                        | ....             | ... |
| 6   | Male   | Z38.0     | Private         | Single liveborn infant, delivered vaginally | No               | 1   |

Gender are crucial demographic factors that significantly influence how severe a disease becomes. Generally, the risk of a severe illness escalates with advancing age, particularly when patients have comorbidities and declining organ function. This increase in risk directly impacts the patient's Length of Stay (LOS) in the hospital [7], [25], [26]. The collected data indicated that male infants constituted 236 cases (53%), whereas female infants accounted for 209 cases (47%).

Insurance claim data has been utilized in several studies to predict LOS; however, only a limited number of investigations have employed hospital records for LOS prediction [13]. In this study, the insurance attribute represents the type of coverage held by the patient, categorized as government insurance, private insurance, or self-payment [27]. Based on the data utilized in this study, 224 cases (50.3%) were covered by government insurance, 218 cases (49.0%) were covered by private insurance, and only 3 cases (0.7%) were paid out-of-pocket.

Diagnosis refers to the determination of a disease, condition, or injury based on observed signs and reported symptoms. It involves gathering and analyzing patient history, conducting physical examinations, and interpreting medical test results to identify the underlying cause and inform appropriate treatment and prognosis. Beyond the medical field, the concept of diagnosis is also applied in areas such as science, engineering, and business to investigate issues and pinpoint their root causes [2].

A referral procedure is a series of steps for transferring a patient from one healthcare facility to another in order to obtain more specialized treatment. Typically, this process begins at a Primary Healthcare Facility and is continued to a higher-level Referral Healthcare Facility when medically necessary. Patients must first undergo an examination at the primary facility, where a physician will decide whether a referral is required. However, in emergency situations, patients may go directly to the nearest hospital without a referral.

Healthcare referral pathways are generally categorized into medical referrals and general health referrals. In the dataset utilized for this study, referral type is represented by two classifications: health referrals and emergency referrals. Of the 445 recorded inpatient cases involving newborns, only 10 cases were identified as emergency referrals, while 1 case involved a newborn admitted in stable, normal condition. Overall, these data indicate that approximately 2.5% of newborn hospitalization cases originated from referral pathways.

Based on Table 1, LOS refers to the duration of a patient's hospitalization. In this study, LOS is categorized into six groups: Group 0 (0-day stay), Group 1 (1–7 days), Group 2 (8–14 days), Group 3 (15–21 days), Group 4 (22–30 days), and Group 5 (more than 30 days). Figure 3 illustrates the distribution of these length-of-stay categories. The results of the analysis can be explained that the LOS data is not balanced from 6 groups, 66.3% for group 1 (see Figure 1).

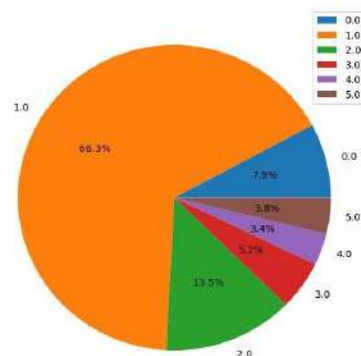


Figure 1. Distribution of LOS Data

As illustrated in Figure 2, the distribution of diagnoses among newborn inpatients includes 145 cases related to immunization-associated health issues, 97 cases involving respiratory and metabolic disorders, 82 cases classified under post-procedural and treatment complications, and 23 cases associated with prematurity. A detailed breakdown of this distribution is presented in Figure 2.

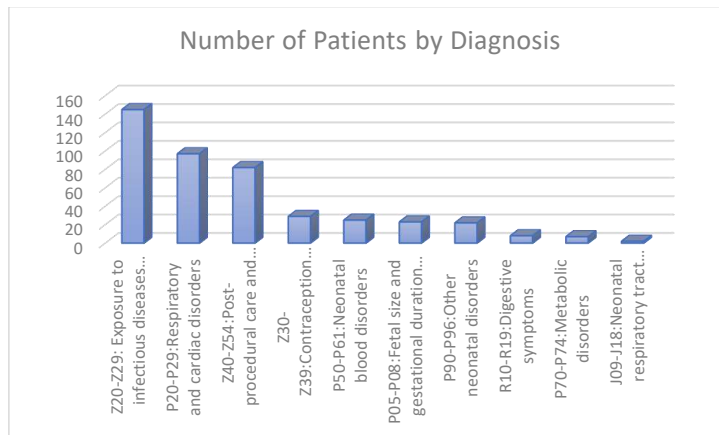


Figure 2. Data Distribution for Diagnosis

### 3.2. Modeling

At this stage, models for regression are developed and optimized using four different algorithms: LR, RFR, SVR, and Artificial Neural Network (ANN), each employing GridSearchCV to identify the best parameters.

- Linear Regression Pipeline: Data preprocessing (numeric features unchanged, categorical features OneHotEncoded) is followed by a Linear Regression model, with GridSearchCV (5-fold CV) used to find the best hyperparameters based on negative mean absolute error.
- Random Forest Regression Pipeline: Data preprocessing is followed by a RandomForestRegressor, with GridSearchCV used to optimize key parameters such as number of estimators, tree depth, and minimum samples for splitting.

Table 2. GridSearchCV-Optimized Hyperparameters for Random Forest Regressor

| Hyperparameter   | Value          |
|------------------|----------------|
| n_estimator      | [50, 100, 200] |
| max_depth        | [None, 10, 20] |
| min_sample_split | [2, 5, 10]     |

In Table 2, the hyperparameters n\_estimators, max\_depth, and min\_samples\_split are presented. The n\_estimators parameter specifies the number of decision trees constructed independently within the ensemble model, which are subsequently combined through a voting mechanism in classification tasks or an averaging process in regression tasks to generate the final prediction.

In this study, three primary hyperparameters were considered to optimize the performance of the ensemble learning model: n\_estimators, max\_depth, and min\_samples\_split. Each of these parameters influences the model's structure and learning behavior in distinct ways.

The n\_estimators parameter determines the number of decision trees constructed independently within the ensemble. These trees are subsequently aggregated through a voting mechanism for classification tasks or an averaging process for regression tasks to produce the final prediction. A higher number of estimators generally enhances model stability and accuracy but also increases computational cost and training time.

The max\_depth parameter defines the maximum allowable depth of each decision tree. This setting regulates the tree's complexity and its ability to capture intricate relationships in the dataset. In this study, three configurations None, 10, and 20 were evaluated. When set to None, the tree expands

without depth limitation, which can lead to overfitting. Restricting the depth to 10 produces a shallower structure that promotes generalization, whereas increasing it to 20 allows for more detailed learning at the expense of higher variance and computational demands.

Finally, the `min_samples_split` parameter specifies the minimum number of samples required to split an internal node. It governs how finely the model partitions the data. The values 2, 5, and 10 were tested to assess their effects on model performance. A smaller value such as 2 permits more frequent splits and results in deeper trees that may overfit, while larger values (e.g., 10) constrain growth, yielding simpler and more stable models that may risk underfitting.

Overall, fine-tuning these hyperparameters is essential to achieving a balance between model bias and variance, thereby improving the ensemble model's predictive capability and generalization performance

- In this pipeline, the SVR model was implemented with a parameter grid search to identify the optimal combination of the kernel type (linear or RBF), the regularization parameter (C), and the epsilon parameter that defines the margin of tolerance.

Table 3. SVR Hyperparameters Tuned Using GridSearchCV

| Hyperparameter | Value             |
|----------------|-------------------|
| Kernel         | ['linear', 'rbf'] |
| C              | [0.1, 1, 10]      |
| Epsilon        | [0.1, 0.2, 0.5]   |

In Table 3, the hyperparameters Kernel, C, and Epsilon are presented. The kernel parameter in Support Vector Regression (SVR) specifies the type of kernel function used to map input data into a higher-dimensional feature space, enabling the model to learn both linear and nonlinear relationships. In this study, two kernel types were utilized: linear and rbf (Radial Basis Function).

The C parameter in Support Vector Regression (SVR) regulates the balance between model complexity and the tolerance for errors in the training data. It serves as a regularization factor that controls how strictly the model fits the training samples. In this study, three values of C were examined: 0.1, 1, and 10.

A lower C value (0.1) applies stronger regularization, allowing a wider margin and greater tolerance for training errors. This configuration tends to produce a smoother and more generalized model but may lead to underfitting if the model becomes too simple. In contrast, a higher C value (10) reduces the regularization effect, forcing the model to minimize training errors and fit the data more closely, which may improve training accuracy but increase the risk of overfitting. The intermediate value  $C = 1$  typically provides a balanced trade-off between these two conditions.

Optimizing the C parameter is essential to achieving an appropriate balance between bias and variance, thereby enhancing the SVR model's ability to generalize effectively to unseen data.

The epsilon parameter in Support Vector Regression (SVR) defines the width of the epsilon-insensitive zone around the regression function, within which errors are not penalized. In other words, it specifies the margin of tolerance where deviations between the predicted and actual values are considered acceptable and do not contribute to the loss function.

In this study, three values of epsilon were examined: 0.1, 0.2, and 0.5. An epsilon value 0.1 creates a narrower margin, making the model more sensitive to small deviations and resulting in a closer fit to the training data, which can increase the risk of overfitting. Conversely, an epsilon value 0.5 establishes a wider margin of tolerance, allowing greater deviations without penalty and promoting a smoother regression function, though it may lead to underfitting if the margin is too broad. The value 0.2 provides a balanced compromise between sensitivity and generalization.

Proper tuning of the epsilon parameter is crucial for controlling the trade-off between model precision and robustness, ensuring that the SVR model captures meaningful trends while maintaining good generalization to unseen data.

- This pipeline employs a Multi-Layer Perceptron (MLP) model, which serves as an implementation of the Artificial Neural Network (ANN). A GridSearchCV procedure was applied to determine the optimal combination of hidden layer sizes, activation functions, and solvers for the model.

Table 4. Hyperparameter ANN Using GridSearchCV

| Hyperparameter    | Value              |
|-------------------|--------------------|
| hidden_layer_size | [(100,), (50, 50)] |
| activation        | ['relu', 'tanh']   |
| solver            | ['adam', 'sgd']    |

Table 4 presents the parameters and corresponding values used in the Multi-Layer Perceptron (MLP) model. The Multilayer Perceptron (MLP) model includes several key hyperparameters that define its architecture and optimization behavior, namely hidden\_layer\_sizes, activation, and solver. These parameters collectively influence the model's learning capacity, convergence speed, and ability to generalize.

The hidden\_layer\_sizes parameter specifies the number of neurons and layers in the network's hidden structure. In this study, two configurations were tested: (100,) and (50, 50). The first configuration, (100,), represents a single hidden layer containing 100 neurons, while (50, 50) denotes a deeper architecture with two hidden layers of 50 neurons each. Increasing the number of layers and neurons generally enhances the model's ability to learn complex relationships, though it may also increase the risk of overfitting and extend training time.

The activation parameter determines the nonlinear transformation applied to the output of each neuron. Two activation functions were considered: 'relu' (Rectified Linear Unit) and 'tanh' (hyperbolic tangent). The 'relu' function is widely used for its computational efficiency and ability to accelerate convergence by mitigating the vanishing gradient problem. In contrast, the 'tanh' function maps input values to a range between -1 and 1, which can be advantageous for normalized data but may lead to slower convergence.

The solver parameter defines the optimization algorithm used to adjust the model's weights during training. Two solvers were employed: 'adam' and 'sgd' (Stochastic Gradient Descent). The 'adam' solver combines the advantages of adaptive learning rate and momentum, making it effective for large or noisy datasets. Meanwhile, the 'sgd' solver updates weights incrementally per batch, providing greater control over learning dynamics but requiring careful tuning of the learning rate to ensure stable convergence.

Overall, tuning these hyperparameters network depth, activation function, and optimization algorithm is essential to achieving a balance between learning efficiency, model complexity, and generalization performance in MLP-based regression models.

### 3.3. Model Evaluation

At this stage, model evaluation was carried out using 5-fold cross-validation along with several performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination ( $R^2$ ), for each trained model—Linear Regression, Random Forest, Support Vector Regression (SVR), and Artificial Neural Network (ANN). These metrics were computed using the cross\_validate function, and their average values were reported for each model. The results provide an overview of model performance in terms of prediction error and the quality of fit to the data.

Cross-validation ensures a more objective evaluation by partitioning the dataset into multiple subsets and testing each subset in turn. Based on the cross-validation results obtained for the four models, a comparative analysis of their performance is presented below. Table 5 presents the results of the experiments.

Table 5. The Results of the Experiments.

| Model                     | MAE   | MSE   | RMSE  | R <sup>2</sup> |
|---------------------------|-------|-------|-------|----------------|
| Linear Regression         | 0.395 | 0.236 | 0.482 | 0.728          |
| Random Forest Regression  | 0.019 | 0.011 | 0.086 | 0.987          |
| Support Vector Regression | 0.370 | 0.260 | 0.500 | 0.731          |
| Artificial Neural Network | 0.253 | 0.107 | 0.324 | 0.868          |

The performance of four regression models—Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN)—was assessed using four statistical metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R<sup>2</sup>). These metrics provide a comprehensive view of each model's predictive accuracy and generalization ability.

As shown in Table 5, Random Forest Regression achieved the best results overall, with the lowest MAE (0.019), MSE (0.011), and RMSE (0.086), alongside the highest R<sup>2</sup> (0.987). This indicates that Random Forest can make highly accurate predictions and effectively capture complex nonlinear relationships in the data. The ANN also demonstrated strong performance, with relatively low error values (MAE = 0.253, RMSE = 0.324) and a high R<sup>2</sup> (0.868). While not as precise as Random Forest, the ANN was able to learn meaningful patterns and generalize well to new data. In contrast, Linear Regression and SVR showed higher errors and lower R<sup>2</sup> values (0.728 and 0.731, respectively), suggesting these models struggle to capture the nonlinear relationships present in the dataset, resulting in less accurate predictions.

Overall, the evaluation indicates that Random Forest Regression is the most reliable approach for this dataset. The ANN offers a solid alternative with strong learning capability, whereas Linear Regression and SVR provide moderate performance.

#### 4. DISCUSSIONS

The experimental results provide a comprehensive comparison of the predictive capabilities of four regression models: Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN). As summarized in Table 5, the evaluation revealed significant differences in accuracy and generalization performance among the models.

Among the tested models, Random Forest Regression demonstrated the highest predictive performance, achieving the lowest errors (MAE = 0.019; MSE = 0.011; RMSE = 0.086) and the highest coefficient of determination (R<sup>2</sup> = 0.987). This suggests that the ensemble learning approach of Random Forest, which combines multiple decision trees through averaging, effectively reduces variance and enhances model robustness. Its ability to capture nonlinear relationships and complex interactions among features makes it particularly suitable for this dataset.

The Artificial Neural Network (ANN) also performed well, showing moderate error values (MAE = 0.253; RMSE = 0.324) and a high R<sup>2</sup> (0.868). The multilayer structure of ANN allows it to model nonlinear relationships in the data, providing better generalization than linear models. However, its accuracy is slightly lower than Random Forest, likely due to sensitivity to hyperparameter settings, training iterations, or potential overfitting in certain configurations.

In contrast, Linear Regression and SVR produced higher errors and lower R<sup>2</sup> scores (0.728 and 0.731, respectively). Linear Regression, constrained by its linearity assumption, struggled to model the

**Commented [A6]:** Penulisan hasil sebaiknya diurutkan berdasarkan urutan tahapan pada bagian Metode, dan harus memuat hasil secara keseluruhan dari penelitian. Penomoran gambar dan tabel harus dirujuk pada teks. Mohon rapikan Tabel 1. Tambahkan tabel "Feature Importance" untuk Random Forest.

complex relationships in the dataset. SVR performance was likely affected by the choice of kernel and regularization settings, which are critical for capturing nonlinear patterns effectively.

The model exhibits considerable potential for further refinement; nevertheless, it requires evaluation by domain experts or end-users, as well as testing on new real-world scenarios to ensure its robustness and applicability.

## 5. CONCLUSION

Based on the comparison of Linear Regression, Random Forest Regression, Support Vector Regression, and Artificial Neural Network algorithms for developing a model to predict the length of stay (LOS) of newborns, the Random Forest Regression method demonstrated the highest performance. The study results indicate that both medical and non-medical factors—such as initial medical condition, primary diagnosis, age, gender, and referral type affect the length of hospitalization for newborn patients. Random Forest Regression showed the best performance, even when applied to a relatively small dataset.

A limitation of this study is the small size of the dataset, which led to suboptimal performance for some of the methods used. Therefore, future research is recommended to incorporate a larger dataset and to explore additional variables, including other medical factors such as birth weight, gestational age, and parental history, as well as non-medical factors such as hospital facilities and geographic location.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the research, authorship, or publication of this article. All stages of the study, including data analysis and manuscript preparation, were conducted objectively without any financial, professional, or institutional influence.

## REFERENCES

- [1] WHO, "Newborn Mortality," WHO. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021>
- [2] Y. G. Robi and T. M. Sitote, "Neonatal Disease Prediction Using Machine Learning Techniques," *J. Healthc. Eng.*, vol. 2023, no. 1, pp. 1–16, 2023, doi: 10.1155/2023/3567194.
- [3] G. Almeida, F. B. Correia, and A. R. Borges, "Hospital Length-of-Stay Prediction Using Machine Learning Algorithms — A Literature Review," *Applied Sci.*, vol. 2024, no. 14, pp. 1–20, 2024, doi: <https://www.mdpi.com/2076-3417/14/22/10523>.
- [4] L. Garg *et al.*, "Characterising Hospital Admission Patterns and Length of Stay in the Emergency Department at Mater Dei Hospital Malta," *Preprints*, vol. 1, no. February, pp. 28–30, 2023, doi: 10.20944/preprints202302.0315.v1.
- [5] K. Stone, R. Zwigelaar, P. Jones, N. Mac Parthaláin, and D. Yoon, "A Systematic Review of The Prediction of Hospital Length of Stay: Towards a Unified Framework," *PLOS Digit. Heal.*, vol. 1, no. 4, pp. 1–39, 2022, doi: 10.1371/journal.pdig.0000017.
- [6] M. A. Shbool *et al.*, "Machine Learning Approaches to Predict Patient's Length of Stay in Emergency Department," *Appl. Comput. Intell. Soft Comput.*, vol. 2023, no. 1, pp. 1–13, 2023, doi: 10.1155/2023/8063846.
- [7] S. G. Gurazada, S. Gao, F. Burstein, and P. Buntine, "Predicting Patient Length of Stay in Australian Emergency Departments Using Data Mining," *Sensors*, vol. 22, no. 13, pp. 1–15, 2022, doi: <https://www.mdpi.com/1424-8220/22/13/4968>.
- [8] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, and A. Duclos, "The Prediction of Hospital Length of Stay using Unstructured Data," *BMC Med. Inform. Decis. Mak.*, vol. 4, pp. 1–9, 2021, doi: 10.1186/s12911-021-01722-4.
- [9] S. Barnes, E. Hamrock, M. Toerper, S. Siddiqui, and S. Levin, "Real-Time Prediction of Inpatient Length of Stay for Discharge Prioritization," *Res. Appl.*, vol. 23, pp. 2–10, 2016, doi: 10.1093/jamia/ocv106.

**Commented [A7]:** Pada bagian Discussion sudah dituliskan pembahasan sesuai pendapat penulis mengenai hasil penelitiannya yang telah dilakukan secara detail. Tambahkan perbandingan dengan penelitian sebelumnya. Tambahkan dampak atau urgensi dari penelitian ini dengan lebih jelas.

**Commented [A8]:** Bab kesimpulan sudah merangkum isi dari hasil diskusi dan pembahasan dan terdapat saran penelitian lanjutan. Tambahkan kalimat eksplisit tentang kontribusi penelitian ke ilmu komputer/informatika.

**Commented [A9]:** Jumlah lebih dari 25 rujukan 80% dari jurnal/conference terbitan 5 tahun terakhir dan sudah menggunakan style IEEE. Mohon perbaiki metadata pada Mendeley atau Zotero, atau EndNote, ada beberapa referensi DOI belum lengkap, nama jurnal tidak lengkap. Saran untuk referensi yang terbit lebih dari 5 tahun mohon dapat diupdate dengan referensi yang terbaru.

- [10] X. Zeng, "Length of Stay Prediction Model of Indoor Patients Based on Light Gradient Boosting Machine," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–14, 2022.
- [11] L. Chen and H. B. Klasky, "Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients with Sepsis : A Comparative Study," *SoutheastCon 2022*, vol. 2022, pp. 302–309, 2022.
- [12] J. Wu *et al.*, "Development of A Scoring Tool for Predicting Prolonged Length of Hospital Stay in Peritoneal Dialysis Patients Through Data Mining," *Ann. Transl. Med.*, vol. 8, no. 21, pp. 1–15, 2020, doi: 10.21037/atm-20-1006.
- [13] J. An, M. Jung, S. Ryu, Y. Choi, and J. Kim, "Analysis of Length of Stay for Patients Admitted to Korean Hospitals Based on The Korean National Health Insurance Service Database," *Informatics Med. Unlocked*, vol. 37, no. October 2022, p. 101178, 2023, doi: 10.1016/j.imu.2023.101178.
- [14] A. I. Taloba, R. M. A. El-aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning," *J. Healthc. Eng.*, vol. 2022, no. 1, pp. 1–10, 2022.
- [15] S. Melhem, A. Al-Aiad, and M. S. Al-Ayyad, "Patient Care Classification Using Machine Learning Techniques," in *International Conference on Information and Communication Systems (ICICS)*, 2021, pp. 1–6. doi: 10.1109/ICICS52457.2021.9464582.
- [16] F. Chollet, *Deep Learning with Python*. United States: Manning Publications Co, 2018. doi: 10.4018/978-1-7998-3095-5.ch003.
- [17] A. Masbakhah, U. Sa'adah, and M. Muslikh, "Heart Disease Classification Using Random Forest and Fox Algorithm as Hyperparameter Tuning," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 4, pp. 964–976, 2025, doi: 10.35882/jeeemi.v7i4.932.
- [18] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, no. April 2015. 2015. doi: 10.1007/978-1-4302-5990-9.
- [19] H. Huang, R. Jia, J. Liang, J. Dang, and Z. Wang, "Wind Power Deterministic Prediction and Uncertainty Quantification Based on Interval Estimation," *J. Sol. Energy Eng. Trans. ASME*, vol. 143, no. 6, pp. 1–12, 2021, doi: 10.1115/1.4051430.
- [20] L. Costa *et al.*, "Multilayer Perceptron," in *Introduction to Computational Intelligence*, no. January, IEEE, 2023, pp. 1–6. doi: 10.1145/3090354.3090427.
- [21] H. Yin, "Enhancing Directional Accuracy in Stock Closing Price Value Prediction Using a Direction-Integrated MSE Loss Function," in *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, 2023, pp. 119–126. doi: 10.5220/0012810200003885.
- [22] T. O. Hodson, "Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Tot," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
- [23] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting Health Care Costs Using Evidence Regression," in *MDPI*, 2019, pp. 1–13. doi: 10.3390/proceedings2019031074.
- [24] D. Sivasundaralingam and N. Thasavaran, "Predicting Inpatient Bed Demand Using Machine Learning," in *YOUNG MEMBERS' TECHNICAL CONFERENCE 2024*, 2023, pp. 150–157. [Online]. Available: [https://www.researchgate.net/publication/391584272\\_Predicting\\_Inpatient\\_Bed\\_Demand\\_Using\\_Machine\\_Learning](https://www.researchgate.net/publication/391584272_Predicting_Inpatient_Bed_Demand_Using_Machine_Learning)
- [25] S. Al-Gahtani and M. M. Shoukri, "Analysis of Length of Stay ( LOS) Data from the Medical Records of Tertiary Care Hospital in Saudi Arabia for Five Diagnosis Related Groups : Application of Cox Prediction Model," *Open J. Stat.*, vol. 11, pp. 99–112, 2021, doi: 10.4236/ojs.2021.111005.
- [26] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-nejad, "Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation," *JMIR Med. INFORMATICS*, vol. 9, no. 5, pp. 1–23, 2021, doi: 10.2196/21347.
- [27] D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon, "Predicting Inpatient Flow at A Major Hospital using Interpretable Analytics," *Manuf. Serv. Oper. Manag.*, vol. 2020, no. 1, pp. 1–42,

2020, doi: <https://doi.org/10.1101/2020.05.12.20098848>.

#### SUBMISSION CHECK-LIST (MUST BE INCLUDED IN ARTICLE FILE)

All items must be YES before submitting to Jurnal Teknik Informatika (JUTIF) in order to make sure a manuscript well-prepared within a high standard. Please put X in column Yes after checking each item

| General Format          |                                                                                                                                                           |     |    |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|----|
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | I have prepared the manuscript using "JUTIF-Template.docx"                                                                                                | X   |    |
| 2                       | The manuscript is written in single or double column using Times New Roman 11 pt font                                                                     | X   |    |
| 3                       | Title is matching with the focus of the manuscript                                                                                                        | X   |    |
| 4                       | I have complete authors information:                                                                                                                      | X   |    |
|                         | a. Affiliation of all authors                                                                                                                             | X   |    |
|                         | b. Indicate one of the authors is the corresponding author                                                                                                | X   |    |
|                         | c. Email of corresponding author                                                                                                                          | X   |    |
| 5                       | The length of the manuscript is not less than 3500 words (excluding abstract, author information and reference list)                                      | X   |    |
| 6                       | The manuscript contains the core contents: introduction with literature review, method, results, discussion, conclusion, conflict of interest, references | X   |    |
| 7                       | Contains maximum 2 level of sections                                                                                                                      | X   |    |
|                         |                                                                                                                                                           |     |    |
| Abstract                |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | The abstract is written between 150 – 250 words                                                                                                           | X   |    |
| 2                       | contains the purpose                                                                                                                                      | X   |    |
| 3                       | contains the method / research approach                                                                                                                   | X   |    |
| 4                       | contains the results / conclusion                                                                                                                         | X   |    |
| 5                       | contains novelty/originality/important findings/impact                                                                                                    | X   |    |
| 6                       | The abstract is completed with keywords                                                                                                                   | X   |    |
|                         |                                                                                                                                                           | X   |    |
| Figures (if applicable) |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | All figures are clear and readable (image, text and legend) with high resolution                                                                          | X   |    |
| 2                       | All figures are in formal style, without redundant title                                                                                                  | X   |    |
| 3                       | All information in the figure are in English and all decimal written in international standard, using point (.) not comma (,)                             | X   |    |
| 4                       | All figures have captions (at the bottom) with consecutive numbers                                                                                        | X   |    |
| 5                       | All figures are cited in the text using consistent citation style                                                                                         | X   |    |
| 6                       | Never citing figure using below and above, but using the number of figure                                                                                 | X   |    |
| Tables (if applicable)  |                                                                                                                                                           |     |    |
| No                      | Item                                                                                                                                                      | Yes | No |
| 1                       | All texts in tables are clear and readable                                                                                                                | X   |    |
| 2                       | All tables are in formal style                                                                                                                            | X   |    |

|                                  |                                                                                                                              |            |           |
|----------------------------------|------------------------------------------------------------------------------------------------------------------------------|------------|-----------|
| 3                                | All information in the table are in English and all decimal written in international standard, using point (.) not comma (,) | X          |           |
| 4                                | All tables are captioned on top with consecutive numbers                                                                     | X          |           |
| 5                                | All tables are cited in the text using consistent citation                                                                   | X          |           |
| 6                                | Never citing table using below and above, but using the number of table                                                      | X          |           |
| <b>Equations (if applicable)</b> |                                                                                                                              |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                  | <b>Yes</b> | <b>No</b> |
| 1                                | All equations are written using editor tool (editable), not a cropped image                                                  | X          |           |
| 2                                | All equations are captioned on top with consecutive numbers                                                                  | X          |           |
| 3                                | All equations are cited in the text using consistent citation                                                                | X          |           |
| 4                                | Never citing equation using below and above, but using the number of equation                                                | X          |           |
| <b>References</b>                |                                                                                                                              |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                  | <b>Yes</b> | <b>No</b> |
| 1                                | Reference list and citation consistently follows the IEEE style                                                              | X          |           |
| 2                                | Journal names in the reference list are not in abbreviated format, but it should be in full name format                      | X          |           |
| 3                                | All references are cited in the text                                                                                         | X          |           |
| 4                                | Citation in the text follows general consistent citation rules                                                               | X          |           |
| 5                                | Paper cites at least 25 references                                                                                           | X          |           |
| 6                                | Book sources are not more than 20% of the reference list                                                                     | X          |           |
| 7                                | 80% of references should come from journals/conferences                                                                      | X          |           |
| 8                                | If references come from journals, it should mention name of journal, volume, no, page, and DOI.                              | X          |           |
| 9                                | Self-citation in the list is not more than 3                                                                                 | X          |           |
| 10                               | Write all authors in the reference list unless authors more than 7 (Write the first 6, then et al.)                          | X          |           |

**Development and Comparative Evaluation of Machine Learning Models using Clinically Relevant Features for Predicting Newborn Patients' Length of Stay****Gandung Triyono<sup>\*1</sup>, Billy Marentek<sup>2</sup>, Mohammad Syafrullah<sup>3</sup>**<sup>1</sup>Information System, Faculty of Information Technology, Budi Luhur University, Indonesia<sup>2,3</sup>Master of Computer Science, Faculty of Information Technology, Budi Luhur University, IndonesiaEmail: <sup>1</sup>gandung.triyono@budiluhur.ac.id

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025

Phone Number : 081382220206

**Abstract**

The Length of Stay (LOS) of newborns is a crucial indicator for healthcare management and hospital resource allocation. However, prior research has yet to systematically compare machine learning models for newborn LOS prediction using clinically pertinent features in developing-country hospital contexts, creating an important methodological and contextual gap. Accurate prediction of LOS is urgently needed to support timely clinical decision-making and prevent overcrowding, inefficiencies, and unnecessary healthcare costs. This study aims to identify factors influencing LOS and develop a predictive model for newborn LOS using several machine learning algorithms. A comparison was conducted among Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Networks (ANN). The dataset consisted of medical records of newborn patients from three private hospitals in Indonesia. The research included data collection and understanding, data preprocessing, modeling, and evaluation. Experimental results show that Random Forest Regression achieved the best predictive performance, with MAE = 0.019, MSE = 0.011, RMSE = 0.086, and  $R^2 = 0.987$ . Feature importance analysis revealed that gender, referral source, insurance type, and diagnosis were the most influential predictors of LOS. This study contributes to the advancement of machine learning applications in healthcare data analytics and provides evidence-based insights to support neonatal care planning and hospital resource optimization.

**Keywords :** *length of stay, machine learning, newborns, prediction, Random Forest;***This work is an open access article licensed under a Creative Commons Attribution 4.0 International License.****1. INTRODUCTION**

Newborn health serves as a key indicator for assessing the quality of a country's healthcare services. Infant mortality and morbidity rates reflect not only the health status of mothers and children but also the overall effectiveness of the healthcare system [1]. As reported by Statistics Indonesia in 2020, although Indonesia has experienced a decline in infant mortality over the past few decades, regional disparities remain significant, indicating inequality in access to and quality of healthcare services. Certain regions, such as Maluku and Papua, continue to record higher infant mortality rates compared to the national average, while areas like Java are approaching the *Sustainable Development Goals* (SDGs) target of 12 deaths per 1,000 live births.

Accurately predicting the LOS of newborns holds significant implications for hospitals in terms of resource management and cost efficiency[2], [3]. Reliable prediction models enable healthcare providers to design more effective care strategies, optimize hospital bed utilization, and reduce patients' financial burdens [3], [4]. However, traditional approaches such as logistic regression and survival analysis often fail to adequately address the non-linear and complex nature of medical data [5], [6].

One of the critical aspects of neonatal care is the length of hospital stay (LOS), which varies considerably depending on several medical factors such as prematurity, respiratory disorders, infections, and congenital anomalies [7]. Prematurity, for example, is a major contributor to extended

hospitalization due to the incomplete development of vital organs, particularly the lungs and immune system [3]. Similarly, Respiratory Distress Syndrome (RDS) is frequently cited as a primary reason for prolonged admission in neonatal intensive care units [8], [9]. In addition, LOS is also influenced by age, gender, and social determinants [4]. The length of hospital stay for newborns is influenced by multiple interconnected factors. Infant-related conditions such as prematurity, low birth weight, infection, respiratory distress syndrome (RDS), and hyperbilirubinemia often require closer monitoring. Maternal factors, including gestational diabetes, pre-eclampsia, and infection, increase neonatal risks and extend care needs. Delivery-related aspects, such as mode of delivery, complications, and premature rupture of membranes, also shape clinical management. Hospital resources like NICU availability and institutional protocols further determine the duration of observation. Social considerations, including long travel distance and unsafe home environments, may delay discharge despite clinical stability [10].

With the advancement of data-driven technologies, machine learning (ML) has emerged as a promising alternative for predictive analytics in healthcare. Numerous studies have demonstrated its potential in forecasting hospital stay durations. Studies such as Almeida [3] highlight that Random Forest can achieve strong predictive performance when supported by effective feature selection. More broadly, existing research shows that LOS prediction does not favor a universally superior algorithm, as outcomes vary with dataset properties and clinical population differences. Even so, ensemble approaches including Random Forest, XGBoost, and LightGBM are repeatedly shown to deliver the most reliable and accurate results, particularly for modeling non-linear patterns and complex feature interactions in diverse healthcare settings [11].

Overall, studies [6], [7], [12], [13], [14], [15] highlight the application of machine learning for predicting Length of Stay (LOS) and patient flow in hospitals and emergency departments (EDs). Study [6] used patient demographics, ED admission status, and inpatient intake information to predict prolonged LOS (>6 days) and LOS as a regression target, with Gradient Boosting performing best for classification (accuracy ~75%, AUC 75.4%, Brier score 0.181) and Ridge and XGBoost excelling in regression. Study [7] emphasized that tree-based models (Random Forest) and neural networks achieved top performance for LOS categories in ED, though model performance heavily depends on local characteristics. Study [8] employed the SPARCS New York dataset for LOS regression with 31 final features, showing LightGBM with proposed encoding as superior ( $R^2$  0.960, MSE 2.231), highlighting boosting models' effectiveness for large, non-linear, complex datasets, with hospital costs as a key predictor. Study [9] used anonymized sepsis data for LOS classification and regression, with XGBoost excelling in classification (accuracy 73.9–79%) and LightGBM yielding the lowest prediction error (MAE  $\approx$  2.4–3.66 days), demonstrating the strength of boosting for complex event data. Study [10] applied Cox Proportional Hazards models to five major DRGs in Saudi Arabia, showing that older age, comorbidities, severity, and emergency admission significantly increased LOS (C-index 0.70–0.85). Study [11] on ICU patients found mortality prediction based on vital signs highly accurate (AUC 0.85–0.92), while LOS prediction was moderate ( $R^2$  0.45–0.62), with XGBoost and Gradient Boosting consistently outperforming other models. Study [12] predicted daily bed demand using historical inpatient data, with XGBoost and LSTM performing best (MAPE 8–12%) and key features including admission/discharge lag, prior occupancy, day of week, holidays, and patient arrival patterns. Study [13] developed an ANN with XAI on 302,966 ED visits, predicting binary admitted/discharged outcomes with AUROC 0.832, recall 75.7%, and accuracy 75.3%; SHAP/PDP identified age, heart rate, and severe injuries as main predictors, demonstrating relevance for patient flow and bed management. Study [14] analyzed 173,005 ED presentations in Victoria, Australia, with LazyIBK (K-NN) and Random Forest achieving best performance (~74% accuracy, ROC 0.81–0.82), highlighting risk factors for LOS >4 hours such as age  $\geq$  64, time to first doctor, arrival mode, triage category, admission flag, and need for CT scan. Study [15] used NHIS Korea data with XGBoost for nationwide LOS prediction, achieving

~0.7472 accuracy and AUC 0.88, with significant features including age, insurance deductible ratio, diagnoses, number of doctors and beds, cholesterol, BMI, and admission month.

Collectively, these studies confirm that machine learning—particularly boosting methods (XGBoost, Gradient Boosting, LightGBM), Random Forest, and neural networks—is effective for predicting LOS and patient flow, with strong evaluation metrics (accuracy 74–83%, AUC 0.75–0.92,  $R^2$  0.45–0.96, low error), and that feature selection, preprocessing, and model interpretability are crucial for clinical implementation, resource management, and bed planning.

Although prior studies have explored machine learning for predicting hospital length of stay, existing work has primarily focused on general patient populations, intensive care settings, or datasets from high-income countries. Current literature lacks a systematic development and comparative evaluation of multiple machine learning algorithms specifically for predicting newborn LOS using clinically relevant variables within the context of hospitals in developing nations. This gap highlights the absence of models that consider regional clinical practices, resource limitations, and population-specific characteristics, thereby creating a significant methodological and contextual void that remains unaddressed.

This study employs several machine learning algorithms Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Artificial Neural Networks (ANN) to predict neonatal LOS, each selected for its ability to manage different levels of data complexity. The urgency of developing an accurate LOS prediction model arises from the fact that imprecise estimations can disrupt neonatal care planning, reduce bed turnover efficiency, and increase operational burdens on healthcare facilities. LR serves as a baseline for capturing linear relationships, RFR models non-linear interactions through ensemble trees, SVR handles complex patterns via kernel functions, and ANN learns high-dimensional representations to enhance predictive precision. Model performance is evaluated using MAE, MSE, RMSE, and  $R^2$ , offering a comprehensive assessment of error magnitude, variance penalization, robustness to outliers, and explanatory power. This study aims to identify key determinants of neonatal LOS and develop models that outperform conventional approaches by incorporating medical, demographic, and socio-economic factors. Overall, the results support more reliable predictive modeling and improved hospital resource planning for neonatal care.

## 2. METHOD

To ensure consistency throughout the research process, this study was structured into several stages that align with its objectives. The main phases of the research include data collection, data preprocessing, modeling, and model evaluation. Each of these stages is described in detail in the following subsections [16]. An overview of the entire research process is illustrated in Figure 1.

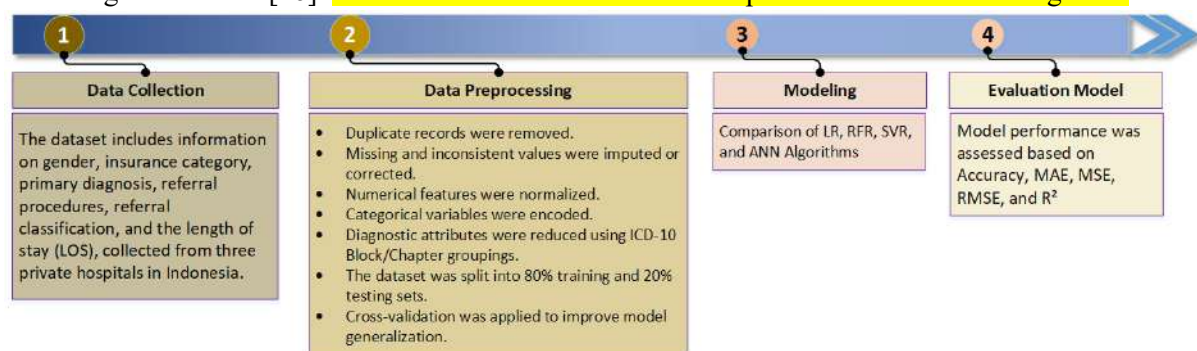


Figure 1. Workflow of LOS Prediction Model Development

### 2.1. Data Collection

The first step of this research involves understanding the business requirements and primary objectives of the study. Hospitals often face challenges in managing bed capacity and medical resources, particularly for newborn patients who require specialized care. Therefore, this study aims to develop a predictive model to estimate the length of stay (LOS) based on both medical and non-medical factors. By utilizing such a model, hospitals can plan resource allocation more effectively and enhance service efficiency.

The study uses data from newborn patients hospitalized in selected hospitals, encompassing clinical and demographic variables such as infant initial diagnosis, medical interventions, and other relevant medical and non-medical factors influencing the duration of hospitalization.

The study population consists of all newborns admitted to the three participating hospitals during the study period. Inclusion criteria encompass newborns with complete medical records who were hospitalized throughout the year 2023. Newborns with incomplete medical records or those referred to other healthcare facilities before completing treatment were excluded from the analysis.

## 2.2. Data Preprocessing

This stage involves several data preprocessing steps to ensure the dataset is in optimal condition before being used for machine learning modeling [17]:

The data preprocessing stage includes several key steps to ensure dataset quality and suitability for model development. Data deduplication is first performed to eliminate duplicate entries and preserve dataset integrity. This is followed by data cleansing, which involves handling missing values through mean or median imputation and removing inconsistent or extreme observations. Data transformation is then applied by normalizing numerical features and encoding categorical variables to ensure proper formatting for machine learning algorithms.

Dimensionality reduction is conducted by utilizing ICD-10 diagnostic codes and extracting their corresponding Block or Chapter levels to produce more compact and meaningful feature representations. Diseases were grouped according to physicians' diagnoses following the ICD-10 system to determine the range of health issues that newborn patients may face.

Finally, the dataset is divided into an 80% training set and a 20% testing set, with cross-validation techniques employed to enhance model generalization and ensure robust performance evaluation.

## 2.3. Modeling

Based on the studies that have been conducted, this study uses four machine learning algorithms to build a prediction model for the length of hospitalization, namely LR, RFR, SVR and ANN.

### 1. Linear Regression

As a supervised machine learning technique, linear regression is trained on labeled datasets to construct the best-fit linear function. This optimized function aims to capture the linear relationship the idea that the output shifts uniformly (at a constant rate) whenever the input changes and is subsequently employed for forecasting values on unseen data. The relationship itself is fundamentally depicted by a straight line [18], [19]. The model of linear regression can be expressed in formula (1).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

where Y represents the dependent variable, which corresponds to the predicted output, while X denotes the independent variable or predictor.  $\beta_0$  refers to the intercept (constant term),  $\beta_1$  is the regression coefficient associated with X, and  $\varepsilon$  represents the error term.

In simple linear regression, the regression coefficients  $\beta_0$  and  $\beta_1$  can be calculated using the Ordinary Least Squares (OLS) method with the formula (2) and (3):

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad (3)$$

where  $X_i$  and  $Y_i$  represent the dependent and independent variables, respectively.  $\bar{X}$  denotes the mean value of  $X$ ,  $\bar{Y}$  denotes the mean value of  $Y$ , and  $n$  refers to the total number of data points.

If there is more than one independent variable, then the model becomes multiple linear regression, method with the formula (4):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (4)$$

where  $Y$  is the dependent variable, namely the predicted output.  $\beta_0$  is an intercept or constant.  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficient for each independent variable.  $X_1, X_2, \dots, X_n$  are an independent variable, namely the predictor.  $\varepsilon$  is an error.

The modeling process using the linear regression algorithm begins by determining the input variables ( $X_1 \dots X_n$ ), such as gender, initial diagnosis, and insurance status, which are assumed to influence the target variable  $Y$ , representing the length of hospitalization. Prior to model training, the dataset is standardized or rescaled to prevent bias arising from differences in feature measurement units. A linear regression model is then applied using the Ordinary Least Squares (OLS) approach to map the relationship between the predictor variables  $X_1 \dots X_n$  and the outcome  $Y$ . During this process, the coefficients  $\beta_0, \beta_1 \dots \beta_n$  are calculated by minimizing the residual error  $\varepsilon$ , enabling the model to generate a regression line that best approximates the actual length of stay experienced by newborn patients.

## 2. Random Forest Regression

Random Forest Regression (RFR) is a powerful ensemble machine learning technique that utilizes a collection of decision trees to achieve a more robust and accurate prediction. The central goal of RFR is to prevent the high variance and overfitting issues typical of single decision trees [20].

Random Forest Regression (RFR) enhances prediction accuracy by introducing randomness during the training process. Each decision tree is constructed using a bootstrap sample, which is a random subset with replacement from the original data, while only a randomly selected subset of features is considered at each node split. By aggregating and averaging the predictions from these independently built trees, the RFR model improves both generalization and prediction stability. Additionally, Out-of-Bag (OOB) samples data not included in the training of each tree are used as an internal validation mechanism to efficiently assess model performance.

The use of RFR learning ensembles, capable of producing models with low variability and higher accuracy. Eq. (5) represent RFR:

$$R(a) = \sum_{k=1}^N Y_k \sum_{j \in C_k} \theta(j; p_k) = \sum_{k=1}^N \sum_{j \in C_k} Y_k \theta(j; p_k) \quad (5)$$

with  $R(a)$  is the final decision of the classifier,  $Y_k$  average response value ke-k,  $C_k$ ,  $P_k$  is a divisive parameter to divide the region of the decision and  $\theta(j; p_k)$  is a function of the information limit based on the  $P_k$  and  $j$ -index. This equation shows the aggregation of the response values in each region  $k$ , taking into account the data subses and relevant divisor parameters [21].

## 3. Support Vector Regression

Support Vector Regression (SVR) is an extension of SVM designed to address regression problems. In SVM, the primary goal is to find the best hyperplane that separates data classes.

Conversely, in SVR, the primary goal is to find a function that predicts continuous values based on the input, while maintaining the SVM's margin principles. SVR works by building a model that not only accounts for prediction errors but also minimizes model complexity. To do this, SVR uses an epsilon-insensitive loss function, which allows small errors to be ignored if they are within an epsilon distance from the predicted value [22].

The Support Vector Regression (SVR) modeling process consists of several key steps. First, a mathematical model is constructed to map input variables, such as diagnosis and gender, to the target variable, namely the length of stay. A tolerance margin is then established to define the acceptable range of prediction errors that do not incur penalties. Errors falling within this range are disregarded to reduce overfitting, while those exceeding the tolerance threshold receive larger penalties. The model is subsequently optimized to maintain a balance between simplicity and predictive accuracy. In cases where the relationships between variables are non-linear, kernel functions are applied to transform the data into a higher-dimensional space, thereby facilitating better pattern recognition. Support vectors—data points located on or beyond the tolerance boundaries—are then identified, as they play a crucial role in determining the structure of the regression function. Finally, the trained SVR model is used to predict the length of stay for new patient data.

SVR maps a low-dimensional input space into a high-dimensional feature space using a nonlinear mapping and then performs a linear regression in the high-dimensional feature space. Given the input dataset  $D = \{(x_i, y_i) : i = 1 \cdots N\}$ , the wind power prediction of SVR method can be given as  $y = f(x) = \omega\phi(x) + bx$ , where  $w$  is the weight,  $b$  is the parameter of bias, and  $\phi(x)$  represents the high-dimensional space. To obtain parameters  $w$  and  $b$ , the prediction of SVR model with structure risk minimization principle can be transformed to a mathematical optimization problem and solved using Lagrange multiplier [23]. The optimization problem is presented as follows (6):

$$f(X, Y_i, Y_i^*) = \sum_{i=1}^N (Y_i - Y_i^*) K(X, X_i) + b \quad (6)$$

Here,  $Y_i$  and  $Y_i^*$  are the Lagrange multipliers, and  $K(X, X_i)$  is the kernel function. The radial basis function (RBF) was used as the SVR kernel, with key parameters kernel coefficient, regularization term, and epsilon margin tuned empirically through multiple trials.

#### 4. Artificial Neural Network

An Artificial Neural Network (ANN) is a computational model designed to approximate complex non-linear mappings through interconnected layers of neurons. The network consists of an input layer, one or more hidden layers, and an output layer. Learning occurs through two main phases: forward propagation and backpropagation [24].

##### 1) Forward Propagation

Each neuron in the hidden layer receives a linear combination of all input variables. The net input for hidden neuron  $j$  is calculated as (7):

$$Z_{tnj} = V_{0j} + \sum_{i=1}^n X_i V_{ij} X_i V_{ij} \quad (7)$$

The activation output is obtained using the sigmoid function (8):

$$Z_j = f(Z_{tnj}), f(x) = \frac{1}{1+e^{-x}} \quad (8)$$

At the output layer, the same computation is applied (9) and (10):

$$Y_{ink} = W_{0k} + \sum_{j=1}^p Z_j W_{jk} \quad (9)$$

$$Y_k = f(Y_{ink}) \quad (10)$$

The output  $Y_k$  is then compared with the target value  $T_k$  to compute the error.

## 2) Backpropagation

Backpropagation is used to update the weights based on the gradient of the error. The output-layer error term is defined as (11):

$$\Delta_k = (T_k - Y_k) f'(Y_{ink}) \quad (11)$$

with the sigmoid derivative (12):

$$f'(x) = f(x)[1 - f(x)] \quad (12)$$

Weight correction between the hidden and output layers is (13):

$$\Delta W_{jk} = \alpha \Delta_k Z_j \quad (13)$$

The error is propagated to the hidden layer (14) and (15):

$$\Delta_{inj} = \sum_{k=1}^m \Delta_k W_{jk} \quad (14)$$

$$\Delta_j = \Delta_{inj} f'(Z_{inj}) \quad (15)$$

Correction of weights between the input and hidden layers is computed as (16):

$$\Delta V_{ij} = \alpha \Delta_j X_i \quad (16)$$

All weights are updated iteratively following (17) and (18):

$$W_{jk}^{new} = W_{jk}^{old} + \Delta W_{jk} \quad (17)$$

$$V_{ij}^{new} = V_{ij}^{old} + \Delta V_{ij} \quad (18)$$

## 3) Testing Phase

During testing, only the forward propagation stage is executed. For binary classification, the final output is assigned as (19):

$$\hat{Y} = \begin{cases} 1, & Y_k \geq 0.5 \\ 0, & Y_k < 0.5 \end{cases} \quad (19)$$

## 2.4 Evaluation Model

After modeling, the next step is to test the model's performance using various evaluation metrics, namely MSE, RMSE, MAE, and  $R^2$ . This testing is done to determine which model provides the most accurate predictions.

Mean Squared Error (MSE) is an equation that measures the average squared error between actual and predicted values. A smaller MSE value indicates better model performance (predictions are closer to reality) [18], [25], as expressed in the following (20):

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (20)$$

where  $n$  represents the sample size,  $y_i$  represents the true value,  $\hat{y}_i$  denotes the predicted value, and  $(y_i - \hat{y}_i)^2$  denotes the squared difference between the actual and predicted values.

The Root Mean Squared Error (RMSE) is a widely used statistical metric that measures the average magnitude of the prediction errors in a regression model. It is defined as the square root of the mean of the squared differences between the actual and predicted values [18], [26], as expressed in the following (21):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (21)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Mean Absolute Error (MAE) is a commonly used evaluation metric that measures the average magnitude of the errors in a set of predictions, without considering their direction. It is defined as the mean of the absolute differences between the actual and predicted values [18], [26], as expressed in the following (22):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (22)$$

where  $y_i$  represents the actual value,  $\hat{y}_i$  denotes the predicted value, and  $n$  is the total number of observations.

The Coefficient of Determination ( $R^2$ ), commonly referred to as R-squared, is a statistical metric used to evaluate the goodness of fit of a regression model. It represents the proportion of the variance in the actual (observed) data that is explained by the predicted values generated by the model [27], [28]. The mathematical formulation of  $R^2$  is expressed as follows (23):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (23)$$

where  $y_i$  denotes the actual value,  $\hat{y}_i$  represents the predicted value,  $\bar{y}$  is the mean of the actual values, and  $n$  is the total number of observations. The numerator  $\sum (y_i - \hat{y}_i)^2$  represents the Residual Sum of Squares (RSS), while the denominator  $\sum (y_i - \bar{y})^2$  represents the Total Sum of Squares (TSS).

An  $R^2$  value close to 1 indicates that the model explains most of the variability in the data, signifying a strong predictive capability. Conversely, an  $R^2$  value near 0 suggests that the model fails to capture the underlying data pattern. In some cases,  $R^2$  can be negative if the model performs worse than a simple mean-based prediction.

### 3. RESULT

The study results are organized into three subsections: Data Collection, Data Preprocessing, Modeling, and Model Evaluation. The first subsection details the dataset and preprocessing outcomes, the second compares the RL, RFR, SVR, and ANN algorithms, and the third presents the evaluation results.

#### 3.1. Data Collection and Preprocessing

The length of stay (LOS) for newborn infants plays a vital role in planning medical care and managing hospital resource allocation. This study aims to determine the factors influencing neonatal LOS and to develop a machine learning-based prediction model that enhances efficiency and accuracy

in hospital management. The data used in this research were collected from three private hospitals in Indonesia.

The dataset successfully collected comprised 3.244 cases obtained from medical records. The data included inpatient information with attributes such as gender, type of insurance, diagnosis, referral procedure, referral type, and Length of Stay (LOS). Following data cleaning and validation in collaboration with hospital physicians, a refined dataset containing 445 cases was produced. This cleaned dataset was subsequently utilized for model development. **Table 1 provides an illustration of the dataset employed in predicting the length of stay (LOS) of newborn patients.**

Table 1. Example of a Dataset Used for Predicting the Length of Stay (LOS) of Newborns

| No  | Gender | Diagnosis | Insurance Claim | Referral Procedure                          | Referral Patient | LOS |
|-----|--------|-----------|-----------------|---------------------------------------------|------------------|-----|
| 1   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 2   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 0   |
| 3   | Male   | P22.9     | Government      | Umbilical venous catheter                   | Yes              | 2   |
| 4   | Female | P07.2     | Private         | Parenteral nutrition                        | Yes              | 1   |
| 5   | Female | Z23       | Government      | Vaccination NEC                             | Yes              | 5   |
| ... | ...    | ...       | ...             | ...                                         | ...              | ... |
| 6   | Male   | Z38.0     | Private         | Single liveborn infant, delivered vaginally | No               | 1   |

Gender are crucial demographic factors that significantly influence how severe a disease becomes. Generally, the risk of a severe illness escalates with advancing age, particularly when patients have comorbidities and declining organ function. This increase in risk directly impacts the patient's Length of Stay (LOS) in the hospital[7], [15], [29]. The collected data indicated that male infants constituted 236 cases (53%), whereas female infants accounted for 209 cases (47%).

Insurance claim data has been utilized in several studies to predict LOS; however, only a limited number of investigations have employed hospital records for LOS prediction [30]. In this study, the insurance attribute represents the type of coverage held by the patient, categorized as government insurance, private insurance, or self-payment [31]. Based on the data utilized in this study, 224 cases (50.3%) were covered by government insurance, 218 cases (49.0%) were covered by private insurance, and only 3 cases (0.7%) were paid out-of-pocket.

Diagnosis refers to the determination of a disease, condition, or injury based on observed signs and reported symptoms. It involves gathering and analyzing patient history, conducting physical examinations, and interpreting medical test results to identify the underlying cause and inform appropriate treatment and prognosis. Beyond the medical field, the concept of diagnosis is also applied in areas such as science, engineering, and business to investigate issues and pinpoint their root causes [2].

A referral procedure is a series of steps for transferring a patient from one healthcare facility to another in order to obtain more specialized treatment. Typically, this process begins at a Primary Healthcare Facility and is continued to a higher-level Referral Healthcare Facility when medically necessary. Patients must first undergo an examination at the primary facility, where a physician will decide whether a referral is required. However, in emergency situations, patients may go directly to the nearest hospital without a referral.

Healthcare referral pathways are generally categorized into medical referrals and general health referrals. In the dataset utilized for this study, referral type is represented by two classifications: health referrals and emergency referrals. Of the 445 recorded inpatient cases involving newborns, only 10 cases were identified as emergency referrals, while 1 case involved a newborn admitted in stable, normal

condition. Overall, these data indicate that approximately 2.5% of newborn hospitalization cases originated from referral pathways.

Based on Table 1, LOS refers to the duration of a patient's hospitalization. In this study, LOS is categorized into six groups: Group 0 (0-day stay), Group 1 (1–7 days), Group 2 (8–14 days), Group 3 (15–21 days), Group 4 (22–30 days), and Group 5 (more than 30 days). Figure 2 illustrates the distribution of these length-of-stay categories. The results of the analysis can be explained that the LOS data is not balanced from 6 groups, 66.3% for group 1 (see Figure 2).

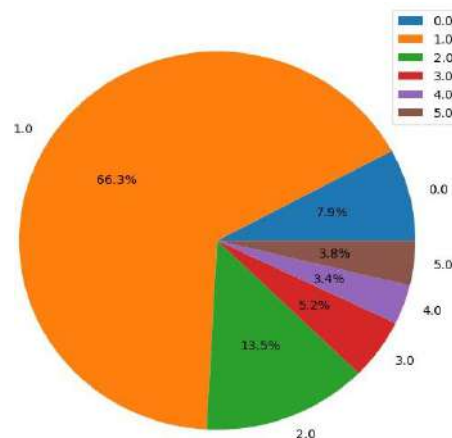


Figure 2. Distribution of LOS Data

Out of the 3,244 collected dataset entries, cases were filtered according to newborn diseases based on physicians' diagnoses. Figure 3 illustrates that, according to ICD-10 classifications, 10 newborn patients were found to require inpatient care. As illustrated in Figure 3, the distribution of diagnoses among newborn inpatients includes 145 cases related to immunization-associated health issues, 97 cases involving respiratory and metabolic disorders, 82 cases classified under post-procedural and treatment complications, and 23 cases associated with prematurity.

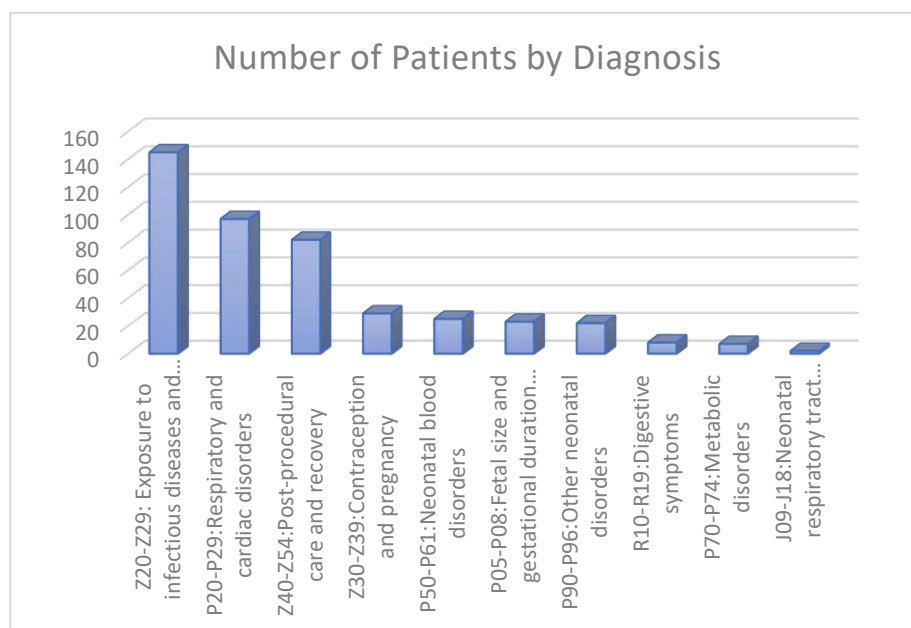


Figure 3. Data Distribution for Diagnosis

### 3.2. Modeling

At this stage, models for regression are developed and optimized using four different algorithms: LR, RFR, SVR, and Artificial Neural Network (ANN), each employing GridSearchCV to identify the best parameters.

- Linear Regression Pipeline: Data preprocessing (numeric features unchanged, categorical features OneHotEncoded) is followed by a Linear Regression model, with GridSearchCV (5-fold CV) used to find the best hyperparameters based on negative mean absolute error.
- Random Forest Regression Pipeline: Data preprocessing is followed by a RandomForestRegressor, with GridSearchCV used to optimize key parameters such as number of estimators, tree depth, and minimum samples for splitting.

Table 2. GridSearchCV-Optimized Hyperparameters for Random Forest Regressor

| Hyperparameter   | Value          |
|------------------|----------------|
| n_estimator      | [50, 100, 200] |
| max_depth        | [None, 10, 20] |
| min_sample_split | [2, 5, 10]     |

In Table 2, the hyperparameters n\_estimators, max\_depth, and min\_samples\_split are presented. The n\_estimators parameter specifies the number of decision trees constructed independently within the ensemble model, which are subsequently combined through a voting mechanism in classification tasks or an averaging process in regression tasks to generate the final prediction.

In this study, three primary hyperparameters were considered to optimize the performance of the ensemble learning model: n\_estimators, max\_depth, and min\_samples\_split. Each of these parameters influences the model's structure and learning behavior in distinct ways.

The n\_estimators parameter determines the number of decision trees constructed independently within the ensemble. These trees are subsequently aggregated through a voting mechanism for classification tasks or an averaging process for regression tasks to produce the final prediction. A higher number of estimators generally enhances model stability and accuracy but also increases computational cost and training time.

The max\_depth parameter defines the maximum allowable depth of each decision tree. This setting regulates the tree's complexity and its ability to capture intricate relationships in the dataset. In this study, three configurations None, 10, and 20 were evaluated. When set to None, the tree expands without depth limitation, which can lead to overfitting. Restricting the depth to 10 produces a shallower structure that promotes generalization, whereas increasing it to 20 allows for more detailed learning at the expense of higher variance and computational demands.

Finally, the min\_samples\_split parameter specifies the minimum number of samples required to split an internal node. It governs how finely the model partitions the data. The values 2, 5, and 10 were tested to assess their effects on model performance. A smaller value such as 2 permits more frequent splits and results in deeper trees that may overfit, while larger values (e.g., 10) constrain growth, yielding simpler and more stable models that may risk underfitting.

Overall, fine-tuning these hyperparameters is essential to achieving a balance between model bias and variance, thereby improving the ensemble model's predictive capability and generalization performance.

In this pipeline, the SVR model was implemented with a parameter grid search to identify the optimal combination of the kernel type (linear or RBF), the regularization parameter (C), and the epsilon parameter that defines the margin of tolerance.

Table 3. SVR Hyperparameters Tuned Using GridSearchCV

| Hyperparameter | Value             |
|----------------|-------------------|
| Kernel         | ['linear', 'rbf'] |
| C              | [0.1, 1, 10]      |
| Epsilon        | [0.1, 0.2, 0.5]   |

In Table 3, the hyperparameters Kernel, C, and Epsilon are presented. The kernel parameter in Support Vector Regression (SVR) specifies the type of kernel function used to map input data into a higher-dimensional feature space, enabling the model to learn both linear and nonlinear relationships. In this study, two kernel types were utilized: linear and rbf (Radial Basis Function).

The C parameter in Support Vector Regression (SVR) regulates the balance between model complexity and the tolerance for errors in the training data. It serves as a regularization factor that controls how strictly the model fits the training samples. In this study, three values of C were examined: 0.1, 1, and 10.

A lower C value (0.1) applies stronger regularization, allowing a wider margin and greater tolerance for training errors. This configuration tends to produce a smoother and more generalized model but may lead to underfitting if the model becomes too simple. In contrast, a higher C value (10) reduces the regularization effect, forcing the model to minimize training errors and fit the data more closely, which may improve training accuracy but increase the risk of overfitting. The intermediate value  $C = 1$  typically provides a balanced trade-off between these two conditions.

Optimizing the C parameter is essential to achieving an appropriate balance between bias and variance, thereby enhancing the SVR model's ability to generalize effectively to unseen data.

The epsilon parameter in Support Vector Regression (SVR) defines the width of the epsilon-insensitive zone around the regression function, within which errors are not penalized. In other words, it specifies the margin of tolerance where deviations between the predicted and actual values are considered acceptable and do not contribute to the loss function.

In this study, three values of epsilon were examined: 0.1, 0.2, and 0.5. An epsilon value 0.1 creates a narrower margin, making the model more sensitive to small deviations and resulting in a closer fit to the training data, which can increase the risk of overfitting. Conversely, an epsilon value 0.5 establishes a wider margin of tolerance, allowing greater deviations without penalty and promoting a smoother regression function, though it may lead to underfitting if the margin is too broad. The value 0.2 provides a balanced compromise between sensitivity and generalization.

Proper tuning of the epsilon parameter is crucial for controlling the trade-off between model precision and robustness, ensuring that the SVR model captures meaningful trends while maintaining good generalization to unseen data.

This pipeline employs a Multi-Layer Perceptron (MLP) model, which serves as an implementation of the Artificial Neural Network (ANN). A GridSearchCV procedure was applied to determine the optimal combination of hidden layer sizes, activation functions, and solvers for the model.

Table 4. Hyperparameter ANN Using GridSearchCV

| Hyperparameter    | Value              |
|-------------------|--------------------|
| hidden_layer_size | [(100,), (50, 50)] |
| activation        | ['relu', 'tanh']   |
| solver            | ['adam', 'sgd']    |

Table 4 presents the parameters and corresponding values used in the Multi-Layer Perceptron (MLP) model. The Multilayer Perceptron (MLP) model includes several key hyperparameters that define its architecture and optimization behavior, namely hidden\_layer\_sizes, activation, and solver.

These parameters collectively influence the model's learning capacity, convergence speed, and ability to generalize.

The `hidden_layer_sizes` parameter specifies the number of neurons and layers in the network's hidden structure. In this study, two configurations were tested: (100,) and (50, 50). The first configuration, (100,), represents a single hidden layer containing 100 neurons, while (50, 50) denotes a deeper architecture with two hidden layers of 50 neurons each. Increasing the number of layers and neurons generally enhances the model's ability to learn complex relationships, though it may also increase the risk of overfitting and extend training time.

The activation parameter determines the nonlinear transformation applied to the output of each neuron. Two activation functions were considered: 'relu' (Rectified Linear Unit) and 'tanh' (hyperbolic tangent). The 'relu' function is widely used for its computational efficiency and ability to accelerate convergence by mitigating the vanishing gradient problem. In contrast, the 'tanh' function maps input values to a range between -1 and 1, which can be advantageous for normalized data but may lead to slower convergence.

The solver parameter defines the optimization algorithm used to adjust the model's weights during training. Two solvers were employed: 'adam' and 'sgd' (Stochastic Gradient Descent). The 'adam' solver combines the advantages of adaptive learning rate and momentum, making it effective for large or noisy datasets. Meanwhile, the 'sgd' solver updates weights incrementally per batch, providing greater control over learning dynamics but requiring careful tuning of the learning rate to ensure stable convergence.

Overall, tuning these hyperparameters network depth, activation function, and optimization algorithm is essential to achieving a balance between learning efficiency, model complexity, and generalization performance in MLP-based regression models.

### 3.3. Model Evaluation

At this stage, model evaluation was carried out using 5-fold cross-validation along with several performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination ( $R^2$ ), for each trained model—Linear Regression, Random Forest, Support Vector Regression (SVR), and Artificial Neural Network (ANN). These metrics were computed using the `cross_validate` function, and their average values were reported for each model. The results provide an overview of model performance in terms of prediction error and the quality of fit to the data.

Cross-validation ensures a more objective evaluation by partitioning the dataset into multiple subsets and testing each subset in turn. Based on the cross-validation results obtained for the four models, a comparative analysis of their performance is presented below. Table 5 presents the results of the experiments.

Table 5. The Results of the Experiments.

| Model                     | MAE   | MSE   | RMSE  | $R^2$ |
|---------------------------|-------|-------|-------|-------|
| Linear Regression         | 0.395 | 0.236 | 0.482 | 0.728 |
| Random Forest Regression  | 0.019 | 0.011 | 0.086 | 0.987 |
| Support Vector Regression | 0.370 | 0.260 | 0.500 | 0.731 |
| Artificial Neural Network | 0.253 | 0.107 | 0.324 | 0.868 |

The performance of four regression models Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN) was assessed using four statistical metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error

(RMSE), and the Coefficient of Determination ( $R^2$ ). These metrics provide a comprehensive view of each model's predictive accuracy and generalization ability.

As shown in Table 5, Random Forest Regression achieved the best results overall, with the lowest MAE (0.019), MSE (0.011), and RMSE (0.086), alongside the highest  $R^2$  (0.987). This indicates that Random Forest can make highly accurate predictions and effectively capture complex nonlinear relationships in the data. The ANN also demonstrated strong performance, with relatively low error values (MAE = 0.253, RMSE = 0.324) and a high  $R^2$  (0.868). While not as precise as Random Forest, the ANN was able to learn meaningful patterns and generalize well to new data. In contrast, Linear Regression and SVR showed higher errors and lower  $R^2$  values (0.728 and 0.731, respectively), suggesting these models struggle to capture the nonlinear relationships present in the dataset, resulting in less accurate predictions.

Overall, the evaluation indicates that Random Forest Regression is the most reliable approach for this dataset. The ANN offers a solid alternative with strong learning capability, whereas Linear Regression and SVR provide moderate performance.

#### 4. DISCUSSIONS

The experimental results provide a comprehensive comparison of the predictive capabilities of four regression models: Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Network (ANN). As summarized in Table 5, the evaluation revealed significant differences in accuracy and generalization performance among the models.

Among the tested models, Random Forest Regression demonstrated the highest predictive performance, achieving the lowest errors (MAE = 0.019; MSE = 0.011; RMSE = 0.086) and the highest coefficient of determination ( $R^2 = 0.987$ ). This suggests that the ensemble learning approach of Random Forest, which combines multiple decision trees through averaging, effectively reduces variance and enhances model robustness. Its ability to capture nonlinear relationships and complex interactions among features makes it particularly suitable for this dataset.

The Artificial Neural Network (ANN) also performed well, showing moderate error values (MAE = 0.253; RMSE = 0.324) and a high  $R^2$  (0.868). The multilayer structure of ANN allows it to model nonlinear relationships in the data, providing better generalization than linear models. However, its accuracy is slightly lower than Random Forest, likely due to sensitivity to hyperparameter settings, training iterations, or potential overfitting in certain configurations.

In contrast, Linear Regression and SVR produced higher errors and lower  $R^2$  scores (0.728 and 0.731, respectively). Linear Regression, constrained by its linearity assumption, struggled to model the complex relationships in the dataset. SVR performance was likely affected by the choice of kernel and regularization settings, which are critical for capturing nonlinear patterns effectively.

**Table 6. Summary of Algorithm Comparisons and Their Performance for LOS Prediction**

| Study | Dataset                       | Algorithms        | Best Performance                      | Features                                                                               |
|-------|-------------------------------|-------------------|---------------------------------------|----------------------------------------------------------------------------------------|
| [6]   | ED admissions                 | Ridge and XGBoost | Accuracy 75%, AUC 75.4%, Brier 0.181; | Used patient demographics, ED admission status, inpatient intake                       |
| [8]   | SPARCS New York (31 features) | LightGBM          | Accuracy 96%, $R^2$ 0.960, MSE 2.231  | Regression; boosting effective for large, non-linear data; hospital cost key predictor |
| [9]   | Sepsis anonymized dataset     | XGBoost, LightGBM | Classification: Accuracy 73.9–79%     | Boosting strong for complex event data                                                 |

|                 |                                     |                              |                                                                    |                                                                                             |
|-----------------|-------------------------------------|------------------------------|--------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| [10]            | 5 major DRGs, Saudi Arabia          | Cox Proportional Hazards     | C-index 0.70–0.85                                                  | LOS influenced by age, comorbidities, severity, emergency admission                         |
| [11]            | ICU patients                        | XGBoost, Gradient Boosting   | Mortality: AUC 0.85–0.92; LOS: $R^2$ 0.45–0.62                     | LOS prediction moderate; boosting superior                                                  |
| [12]            | Daily bed demand                    | XGBoost, LSTM                | MAPE 8–12%                                                         | Admission/discharge lag, prior occupancy, day of week, holidays, arrival patterns           |
| [13]            | 302,966 ED visits                   | ANN + XAI                    | Accuracy 75.3%, Recall 75.7%, AUROC 0.832                          | Age, heart rate, severe injuries                                                            |
| [14]            | 173,005 ED presentations, Australia | LazyIBK (KNN), Random Forest | Accuracy 74%, ROC 0.81–0.82                                        | LOS >4h influenced by age $\geq 64$ , triage, arrival mode, CT scan, admission flag         |
| [15]            | NHIS Korea                          | XGBoost                      | Accuracy 74.72%, AUC 88%                                           | Age, insurance deductible ratio, diagnoses, doctors/beds, BMI, cholesterol, admission month |
| Research Result | Neonatal Data                       | Random Forest Regression     | Accuracy 98.7%, MAE 0.019, MSE 0.011, and RMSE 0.086, $R^2$ 0.987. | Gender, type of insurance, diagnosis, referral procedure, referral type                     |

Table 6 shows that Random Forest Regression achieved the highest performance relative to prior research, though direct comparisons are limited by differences in datasets. The model exhibits considerable potential for further refinement; nevertheless, it requires evaluation by domain experts or end-users, as well as testing on new real-world scenarios to ensure its robustness and applicability.

## 5. CONCLUSION

Based on the comparison of Linear Regression, Random Forest Regression, Support Vector Regression, and Artificial Neural Network algorithms for developing a model to predict the length of stay (LOS) of newborns, the Random Forest Regression method demonstrated the highest performance. The study results indicate that both medical and non-medical factors such as initial medical condition, primary diagnosis, age, gender, and referral type affect the length of hospitalization for newborn patients. Random Forest Regression showed the best performance, even when applied to a relatively small dataset, highlighting its robustness and reliability. These findings contribute to improving predictive accuracy in neonatal LOS estimation and provide valuable insights for healthcare planning and resource allocation.

A limitation of this study is the small size of the dataset, which led to suboptimal performance for some of the methods used. Therefore, future research is recommended to incorporate a larger dataset and to explore additional variables, including other medical factors such as birth weight, gestational age, and parental history, as well as non-medical factors such as hospital facilities and geographic location.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the research, authorship, or publication of this article. All stages of the study, including data analysis and manuscript preparation, were conducted objectively without any financial, professional, or institutional influence.

**REFERENCES**

- [1] WHO, "Newborn Mortality," WHO. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021>
- [2] Y. G. Robi and T. M. Sitote, "Neonatal Disease Prediction Using Machine Learning Techniques," *J. Healthc. Eng.*, vol. 2023, no. 1, pp. 1–16, 2023, doi: 10.1155/2023/3567194.
- [3] G. Almeida, F. B. Correia, and A. R. Borges, "Hospital Length-of-Stay Prediction Using Machine Learning Algorithms — A Literature Review," *Applied Sci.*, vol. 2024, no. 14, pp. 1–20, 2024, doi: <https://www.mdpi.com/2076-3417/14/22/10523>.
- [4] L. Garg *et al.*, "Characterising Hospital Admission Patterns and Length of Stay in the Emergency Department at Mater Dei Hospital Malta," *Preprints*, vol. 1, no. February, pp. 28–30, 2023, doi: 10.20944/preprints202302.0315.v1.
- [5] K. Stone, R. Zwigelaar, P. Jones, N. Mac Parthaláin, and D. Yoon, "A Systematic Review of The Prediction of Hospital Length of Stay: Towards a Unified Framework," *PLOS Digital Heal.*, vol. 1, no. 4, pp. 1–39, 2022, doi: 10.1371/journal.pdig.0000017.
- [6] M. A. Shbool *et al.*, "Machine Learning Approaches to Predict Patient's Length of Stay in Emergency Department," *Appl. Comput. Intell. Soft Comput.*, vol. 2023, no. 1, pp. 1–13, 2023, doi: 10.1155/2023/8063846.
- [7] S. G. Gurazada, S. Gao, F. Burstein, and P. Buntine, "Predicting Patient Length of Stay in Australian Emergency Departments Using Data Mining," *Sensors*, vol. 22, no. 13, pp. 1–15, 2022, doi: <https://www.mdpi.com/1424-8220/22/13/4968>.
- [8] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, and A. Duclos, "The Prediction of Hospital Length of Stay using Unstructured Data," *BMC Med. Inform. Decis. Mak.*, vol. 4, pp. 1–9, 2021, doi: 10.1186/s12911-021-01722-4.
- [9] É. Arnaud *et al.*, "Development and Clinical Interpretation of an Explainable AI Model for Predicting Patient Pathways in the Emergency Department : A Retrospective Study," *Applied Sci.*, vol. 15, no. 2025, pp. 1–22, 2025, doi: <https://doi.org/10.3390/app15158449>.
- [10] Y. Liu *et al.*, "Evaluation of the Need for Intensive Care in Children With Pneumonia : Machine Learning Approach Corresponding," *JMIR Med. INFORMATICS*, vol. 10, no. 1, pp. 1–11, 2022, doi: 10.2196/28934.
- [11] J. Wu *et al.*, "Development of A Scoring Tool for Predicting Prolonged Length of Hospital Stay in Peritoneal Dialysis Patients Through Data Mining," *Ann. Transl. Med.*, vol. 8, no. 21, pp. 1–15, 2020, doi: 10.21037/atm-20-1006.
- [12] A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine Learning-Based Prediction of Hospital Prolonged Length of Stay Admission at Emergency Department : a Gradient Boosting Algorithm Analysis," *Front. Artif. Intell.*, vol. 6, no. 2023, pp. 1–19, 2023, doi: 10.3389/frai.2023.1179226.
- [13] X. Zeng, "Length of Stay Prediction Model of Indoor Patients Based on Light Gradient Boosting Machine," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–14, 2022, doi: <https://doi.org/10.1155/2022/9517029> Research.
- [14] L. Chen and H. B. Klasky, "Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients with Sepsis : A Comparative Study," *SoutheastCon 2022*, vol. 2022, pp. 302–309, 2022, doi: 10.1109/SoutheastCon48659.2022.9764052.
- [15] S. Al-Gahtani and M. M. Shoukri, "Analysis of Length of Stay ( LOS) Data from the Medical Records of Tertiary Care Hospital in Saudi Arabia for Five Diagnosis Related Groups : Application of Cox Prediction Model," *Open J. Stat.*, vol. 11, pp. 99–112, 2021, doi: 10.4236/ojs.2021.111005.
- [16] A. Massahiro *et al.*, "The Evolution of CRISP-DM for Data Science : Methods , Processes and Frameworks," *SBC Rev. Comput. Sci.*, vol. 2023, no. 4, pp. 1–16, 2024, doi: 10.5753/reviews.2024.3757.
- [17] Y. Taules, S. Gros, M. Viladrosa, N. Llorens, S. Solis, and O. Yuguero, "Use of Artificial Intelligence for Reverse Referral Between A Hospital Emergency Department and A Primary Urgent Care Center," *Front. Digit. Heal.*, vol. 7, no. March, pp. 5–10, 2025, doi: 10.3389/fdgth.2025.1546467.
- [18] A. I. Taloba, R. M. A. El-aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and

- Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning,” *J. Healthc. Eng.*, vol. 2022, no. 1, pp. 1–10, 2022, doi: <https://doi.org/10.1155/2022/7969220> Research.
- [19] S. Melhem, A. Al-Aiad, and M. S. Al-Ayyad, “Patient Care Classification Using Machine Learning Techniques,” in *International Conference on Information and Communication Systems (ICICS)*, 2021, pp. 1–6. doi: 10.1109/ICICS52457.2021.9464582.
- [20] F. Chollet, *Deep Learning with Python*. United States: Manning Publications Co, 2018. doi: 10.4018/978-1-7998-3095-5.ch003.
- [21] A. Masbakhah, U. Sa’adah, and M. Muslikh, “Heart Disease Classification Using Random Forest and Fox Algorithm as Hyperparameter Tuning,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 4, pp. 964–976, 2025, doi: 10.35882/jeeemi.v7i4.932.
- [22] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, no. April 2015. 2015. doi: 10.1007/978-1-4302-5990-9.
- [23] H. Huang, R. Jia, J. Liang, J. Dang, and Z. Wang, “Wind Power Deterministic Prediction and Uncertainty Quantification Based on Interval Estimation,” *J. Sol. Energy Eng. Trans. ASME*, vol. 143, no. 6, pp. 1–12, 2021, doi: 10.1115/1.4051430.
- [24] L. Costa *et al.*, “Multilayer Perceptron,” in *Introduction to Computational Intelligence*, no. January, IEEE, 2023, pp. 1–6. doi: 10.1145/3090354.3090427.
- [25] H. Yin, “Enhancing Directional Accuracy in Stock Closing Price Value Prediction Using a Direction-Integrated MSE Loss Function,” in *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, 2023, pp. 119–126. doi: 10.5220/0012810200003885.
- [26] T. O. Hodson, “Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Tot,” *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
- [27] A. Danujan, T. Nithusikan, Athauda, Hearth, and Senanayake, “Predicting Inpatient Bed Demand Using Machine Learning,” in *Young Members’ Technical Conference 2024*, 2023, pp. 150–157. [Online]. Available: [https://www.researchgate.net/publication/391584272\\_Predicting\\_Inpatient\\_Bed\\_Demand\\_Using\\_Machine\\_Learning](https://www.researchgate.net/publication/391584272_Predicting_Inpatient_Bed_Demand_Using_Machine_Learning)
- [28] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, “Predicting Health Care Costs Using Evidence Regression,” in *MDPI*, 2019, pp. 1–13. doi: 10.3390/proceedings2019031074.
- [29] K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-nejad, “Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation,” *JMIR Med. INFORMATICS*, vol. 9, no. 5, pp. 1–23, 2021, doi: 10.2196/21347.
- [30] J. An, M. Jung, S. Ryu, Y. Choi, and J. Kim, “Analysis of Length of Stay for Patients Admitted to Korean Hospitals Based on The Korean National Health Insurance Service Database,” *Informatics Med. Unlocked*, vol. 37, no. October 2022, p. 101178, 2023, doi: 10.1016/j.imu.2023.101178.
- [31] D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon, “Predicting Inpatient Flow at A Major Hospital using Interpretable Analytics,” *Manuf. Serv. Oper. Manag.*, vol. 2020, no. 1, pp. 1–42, 2020, doi: <https://doi.org/10.1101/2020.05.12.20098848>.

### SUBMISSION CHECK-LIST (MUST BE INCLUDED IN ARTICLE FILE)

All items must be YES before submitting to Jurnal Teknik Informatika (JUTIF) in order to make sure a manuscript well-prepared within a high standard. Please put X in column Yes after checking each item

| General Format |                                                            |     |    |
|----------------|------------------------------------------------------------|-----|----|
| No             | Item                                                       | Yes | No |
| 1              | I have prepared the manuscript using “JUTIF-Template.docx” | X   |    |

|                                  |                                                                                                                                                           |            |           |
|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|------------|-----------|
| 2                                | The manuscript is written in single or double column using Times New Roman 11 pt font                                                                     | X          |           |
| 3                                | Title is matching with the focus of the manuscript                                                                                                        | X          |           |
| 4                                | I have complete authors information:                                                                                                                      | X          |           |
|                                  | a. Affiliation of all authors                                                                                                                             | X          |           |
|                                  | b. Indicate one of the authors is the corresponding author                                                                                                | X          |           |
|                                  | c. Email of corresponding author                                                                                                                          | X          |           |
| 5                                | The length of the manuscript is not less than 3500 words (excluding abstract, author information and reference list)                                      | X          |           |
| 6                                | The manuscript contains the core contents: introduction with literature review, method, results, discussion, conclusion, conflict of interest, references | X          |           |
| 7                                | Contains maximum 2 level of sections                                                                                                                      | X          |           |
|                                  |                                                                                                                                                           |            |           |
| <b>Abstract</b>                  |                                                                                                                                                           |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                                               | <b>Yes</b> | <b>No</b> |
| 1                                | The abstract is written between 150 – 250 words                                                                                                           | X          |           |
| 2                                | contains the purpose                                                                                                                                      | X          |           |
| 3                                | contains the method / research approach                                                                                                                   | X          |           |
| 4                                | contains the results / conclusion                                                                                                                         | X          |           |
| 5                                | contains novelty/originality/important findings/impact                                                                                                    | X          |           |
| 6                                | The abstract is completed with keywords                                                                                                                   | X          |           |
|                                  |                                                                                                                                                           | X          |           |
| <b>Figures (if applicable)</b>   |                                                                                                                                                           |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                                               | <b>Yes</b> | <b>No</b> |
| 1                                | All figures are clear and readable (image, text and legend) with high resolution                                                                          | X          |           |
| 2                                | All figures are in formal style, without redundant title                                                                                                  | X          |           |
| 3                                | All information in the figure are in English and all decimal written in international standard, using point (.) not comma (,)                             | X          |           |
| 4                                | All figures have captions (at the bottom) with consecutive numbers                                                                                        | X          |           |
| 5                                | All figures are cited in the text using consistent citation style                                                                                         | X          |           |
| 6                                | Never citing figure using below and above, but using the number of figure                                                                                 | X          |           |
| <b>Tables (if applicable)</b>    |                                                                                                                                                           |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                                               | <b>Yes</b> | <b>No</b> |
| 1                                | All texts in tables are clear and readable                                                                                                                | X          |           |
| 2                                | All tables are in formal style                                                                                                                            | X          |           |
| 3                                | All information in the table are in English and all decimal written in international standard, using point (.) not comma (,)                              | X          |           |
| 4                                | All tables are captioned on top with consecutive numbers                                                                                                  | X          |           |
| 5                                | All tables are cited in the text using consistent citation                                                                                                | X          |           |
| 6                                | Never citing table using below and above, but using the number of table                                                                                   | X          |           |
| <b>Equations (if applicable)</b> |                                                                                                                                                           |            |           |
| <b>No</b>                        | <b>Item</b>                                                                                                                                               | <b>Yes</b> | <b>No</b> |
| 1                                | All equations are written using editor tool (editable), not a cropped image                                                                               | X          |           |
| 2                                | All equations are captioned on top with consecutive numbers                                                                                               | X          |           |
| 3                                | All equations are cited in the text using consistent citation                                                                                             | X          |           |

|                   |                                                                                                         |            |           |
|-------------------|---------------------------------------------------------------------------------------------------------|------------|-----------|
| 4                 | Never citing equation using below and above, but using the number of equation                           | X          |           |
| <b>References</b> |                                                                                                         |            |           |
| <b>No</b>         | <b>Item</b>                                                                                             | <b>Yes</b> | <b>No</b> |
| 1                 | Reference list and citation consistently follows the IEEE style                                         | X          |           |
| 2                 | Journal names in the reference list are not in abbreviated format, but it should be in full name format | X          |           |
| 3                 | All references are cited in the text                                                                    | X          |           |
| 4                 | Citation in the text follows general consistent citation rules                                          | X          |           |
| 5                 | Paper cites at least 25 references                                                                      | X          |           |
| 6                 | Book sources are not more than 20% of the reference list                                                | X          |           |
| 7                 | 80% of references should come from journals/conferences                                                 | X          |           |
| 8                 | If references come from journals, it should mention name of journal, volume, no, page, and DOI.         | X          |           |
| 9                 | Self-citation in the list is not more than 3                                                            | X          |           |
| 10                | Write all authors in the reference list unless authors more than 7 (Write the first 6, then et al.)     | X          |           |



Gandung Triyono <gandung.triyono@budiluhur.ac.id>

---

### [Jutif] Editor Decision

2 pesan

Nur Chasanah <jutif.ft@unsoed.ac.id>

5 Januari 2026 pukul 11.57

Kepada: Gandung Triyono <gandung.triyono@budiluhur.ac.id>

Dear Dr/Mr/Ms. Gandung Triyono,

We are glad to inform you that your submission with the title "Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay" **HAS BEEN ACCEPTED** for publication in Jurnal Teknik Informatika (Jutif), with e-ISSN: 2723-3871 and is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024.

Regarding to the accepted paper, we will send a Letter of Acceptance (LoA) and will publish your paper after the authors paid the publication fee. The amount of publication fee is : **Rp 1.500.000,00 for normal** (will be published ordered by quota) or **Rp 2.500.000 for fasttrack** (by confirmation, will be published in the nearest fasttrack available edition).

Please transfer the publication fee to :

- **Normal / Fasttrack : Bank BNI account number 1982343238 (JUTIF KELOMPOK).**

After completing the payment, please confirm/send proof of transfer via Whatsapp to 0823-2492-4093.

For further information/questions, please contact 0823-2492-4093 .

We encourage you to submit your other research results to Jurnal Teknik Informatika (Jutif) in the future.

Thank you for your contribution.

---

Jurnal Teknik Informatika (Jutif)

---

3 lampiran

 **A-jutif-review-assignment-5410-Article+Manuscript-16898 - Reviewed.docx**  
168K

 **B-jutif-review-assignment-5410-Article+Manuscript-16898\_rev.docx**  
171K

 **C-JUTIF-Template.docx**  
226K

---

**Lasmedi Afuan** <jutif.ft@unsoed.ac.id>

5 Januari 2026 pukul 19.42

Kepada: Gandung Triyono <gandung.triyono@budiluhur.ac.id>

Gandung Triyono:

The editing of your submission, "Development and Comparison of Machine Learning Models for Predicting Newborn Patients' Length of Stay," is complete. We are now sending it to production.

Submission URL: <https://jutif.if.unsoed.ac.id/index.php/jurnal/authorDashboard/submission/5410>

[Kutipan teks disembunyikan]

January 5, 2026

No : 5410/LoA/JUTIF/I/2026  
 About : Letter of Acceptance

To

1. Gandung Triyono, Budi Luhur University, Indonesia
2. Billy Marentek, Budi Luhur University, Indonesia
3. Mohammad Syafrullah, Budi Luhur University, Indonesia

**Jurnal Teknik Informatika (JUTIF)** is a national journal registered to BRIN with P-ISSN: 2723-3863 and e-ISSN: 2723-3871. JUTIF is accredited SINTA 2 based on the Decree of the Director General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia Number 177/E/KPT/2024.

We would like to thank you for submitting scientific article to the **Jurnal Teknik Informatika (JUTIF)** below:

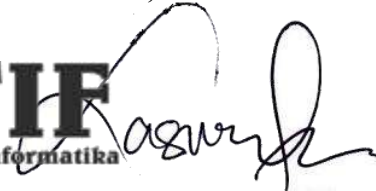
|         |                                                                                                                                                                                                                                       |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Title   | <b>Development and Comparative Evaluation of Machine Learning Models using Clinically Relevant Features for Predicting Newborn Patients' Length of Stay</b>                                                                           |
| Authors | <ol style="list-style-type: none"> <li>1. Gandung Triyono, Budi Luhur University, Indonesia</li> <li>2. Billy Marentek, Budi Luhur University, Indonesia</li> <li>3. Mohammad Syafrullah, Budi Luhur University, Indonesia</li> </ol> |

Based on the review results, the article is **ACCEPTED** for publication in the **Jurnal Teknik Informatika (JUTIF)** Universitas Jenderal Soedirman (UNSOED) **Volume 7 Number 3, June 2026**.

We express our gratitude for your participation and good cooperation for publishing this article.

Chief Editor,


**JUTIF**  
 Jurnal Teknik Informatika



**Dr. Ir. Lasmedi Afuan, S.T., M.Cs., IPM.**  
 NIP. 19850510 200812 1 002

## 11. Bukti Penerbitan Paper





JUTIF

Jurnal Teknik Informatika

p-ISSN : 2723-3863

e-ISSN : 2723-3871

HOME

ABOUT ▾

CURRENT

ARCHIVES

ANNOUNCEMENTS

FAQ

Dashboard 0

Q SEARCH

HOME / ARCHIVES / VOL. 7 NO. 3 (2026): JUTIF VOLUME 7, NUMBER 3, JUNE 2026 / Articles

Development and Comparative Evaluation of Machine Learning Models using Clinically Relevant Features for Predicting Newborn Patients' Length of Stay

Gandung Triyono

Information System, Faculty of Information Technology, Budi Luhur University, Indonesia

Billy Marentek

Master of Computer Science, Faculty of Information Technology, Budi Luhur University, Indonesia

Mohammad Syafrullah

Master of Computer Science, Faculty of Information Technology, Budi Luhur University, Indonesia

DOI: <https://doi.org/10.52436/1.jutif.2026.7.3.5410>

Keywords: length of stay, machine learning, newborns, prediction, Random Forest



DOWNLOAD PDF

MAKE A SUBMISSION

Download Template





</a

ABSTRACT

The Length of Stay (LOS) of newborns is a crucial indicator for healthcare management and hospital resource allocation. However, prior research has yet to systematically compare machine learning models for newborn LOS prediction using clinically pertinent features in developing-country hospital contexts, creating an important methodological and contextual gap. Accurate prediction of LOS is urgently needed to support timely clinical decision-making and prevent overcrowding, inefficiencies, and unnecessary healthcare costs. This study aims to identify factors influencing LOS and develop a predictive model for newborn LOS using several machine learning algorithms. A comparison was conducted among Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Networks (ANN). The dataset consisted of medical records of newborn patients from three private hospitals in Indonesia. The research included data collection and understanding, data preprocessing, modeling, and evaluation. Experimental results show that Random Forest Regression achieved the best predictive performance, with MAE = 0.019, MSE = 0.011, RMSE = 0.086, and  $R^2 = 0.987$ . Feature importance analysis revealed that gender, referral source, insurance type, and diagnosis were the most influential predictors of LOS. This study contributes to the advancement of machine learning applications in healthcare data analytics and provides evidence-based insights to support neonatal care planning and hospital resource optimization.

DOWNLOADS

Download data is not yet available.

REFERENCES

PUBLISHED

2026-06-15

HOW TO CITE

G. Triyono, B. . Marentek, and M. . Syafrullah, "Development and Comparative Evaluation of Machine Learning Models using Clinically Relevant Features for Predicting Newborn Patients' Length of Stay", *J. Tek. Inform. (JUTIF)*, vol. 7, no. 3, pp. 2323–2339, Jun. 2026.

More Citation Formats

ISSUE

Vol. 7 No. 3 (2026): JUTIF  
Volume 7, Number 3, June  
2026

SECTION

Articles

LICENSE

Copyright (c) 2026 Gandung Triyono,  
Billy Marentek, Mohammad Syafrullah

...: ISSN :...

P-ISSN : 2723-3863  
e-ISSN : 2723-3871

...: ABOUT JUTIF :...

Focus and Scope

Editorial Team

Reviewers

Author Guidelines

Peer Review Process

Open Access Policy

Licensing & Copyright

Archiving

Plagiarism Policy

Publication Frequency

Publication Ethics

Article Processing Charge

Generative AI Policy

Indexing

Journal's History

Contact

Article Cited in

Scopus

Tools

WHO, “Newborn Mortality,” WHO. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021>

Y. G. Robi and T. M. Sitote, “Neonatal Disease Prediction Using Machine Learning Techniques,” J. Healthc. Eng., vol. 2023, no. 1, pp. 1–16, 2023, doi: 10.1155/2023/3567194.

G. Almeida, F. B. Correia, and A. R. Borges, “Hospital Length-of-Stay Prediction Using Machine Learning Algorithms — A Literature Review,” Applid Sci., vol. 2024, no. 14, pp. 1–20, 2024, doi: <https://www.mdpi.com/2076-3417/14/22/10523>.

L. Garg et al., “Characterising Hospital Admission Patterns and Length of Stay in the Emergency Department at Mater Dei Hospital Malta,” Preprints, vol. 1, no. February, pp. 28–30, 2023, doi: 10.20944/preprints202302.0315.v1.

K. Stone, R. Zwigelaar, P. Jones, N. Mac Parthaláin, and D. Yoon, “A Systematic Review of The Prediction of Hospital Length of Stay: Towards a Unified Framework,” PLOS Ddigital Heal., vol. 1, no. 4, pp. 1–39, 2022, doi: 10.1371/journal.pdig.0000017.

M. A. Shbool et al., “Machine Learning Approaches to Predict Patient’s Length of Stay in Emergency Department,” Appl. Comput. Intell. Soft Comput., vol. 2023, no. 1, pp. 1–13, 2023, doi: 10.1155/2023/8063846.

S. G. Gurazada, S. Gao, F. Burstein, and P. Buntine, “Predicting Patient Length of Stay in Australian Emergency Departments Using Data Mining,” Sensors, vol. 22, no. 13, pp. 1–15, 2022, doi: <https://www.mdpi.com/1424-8220/22/13/4968>.

J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, and A. Duclos, “The Prediction of Hospital Length of Stay using Unstructured Data,” BMC Med. Inform. Decis. Mak., vol. 4, pp. 1–9, 2021, doi: 10.1186/s12911-021-01722-4.

É. Arnaud et al., “Development and Clinical Interpretation of an Explainable AI Model for Predicting Patient Pathways in the Emergency Department : A Retrospective Study,” Applid Sci., vol. 15, no. 2025, pp. 1–22, 2025, doi: <https://>



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



Free counters!

## INFORMATION

For Readers

For Authors

For Librarians

[doi.org/10.3390/app15158449](https://doi.org/10.3390/app15158449).

Y. Liu et al., “Evaluation of the Need for Intensive Care in Children With Pneumonia : Machine Learning Approach Corresponding,” JMIR Med. INFORMATICS, vol. 10, no. 1, pp. 1–11, 2022, doi: 10.2196/28934.

J. Wu et al., “Development of A Scoring Tool for Predicting Prolonged Length of Hospital Stay in Peritoneal Dialysis Patients Through Data Mining,” Ann. Transl. Med., vol. 8, no. 21, pp. 1–15, 2020, doi: 10.21037/atm-20-1006.

A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, “Machine Learning-Based Prediction of Hospital Prolonged Length of Stay Admission at Emergency Department : a Gradient Boosting Algorithm Analysis,” Front. Artif. Intell., vol. 6, no. 2023, pp. 1–19, 2023, doi: 10.3389/frai.2023.1179226.

X. Zeng, “Length of Stay Prediction Model of Indoor Patients Based on Light Gradient Boosting Machine,” Comput. Intell. Neurosci., vol. 2022, no. 1, pp. 1–14, 2022, doi: <https://doi.org/10.1155/2022/9517029> Research.

L. Chen and H. B. Klasky, “Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients with Sepsis : A Comparative Study,” SoutheastCon 2022, vol. 2022, pp. 302–309, 2022, doi: 10.1109/SoutheastCon48659.2022.9764052.

S. Al-Gahtani and M. M. Shoukri, “Analysis of Length of Stay ( LOS) Data from the Medical Records of Tertiary Care Hospital in Saudi Arabia for Five Diagnosis Related Groups : Application of Cox Prediction Model,” Open J. Stat., vol. 11, pp. 99–112, 2021, doi: 10.4236/ojs.2021.111005.

A. Massahiro et al., “The Evolution of CRISP-DM for Data Science : Methods , Processes and Frameworks,” SBC Rev. Comput. Sci., vol. 2023, no. 4, pp. 1–16, 2024, doi: 10.5753/reviews.2024.3757.

Y. Taules, S. Gros, M. Viladrosa, N. Llorens, S. Solis, and O. Yuguero, “Use of Artificial Intelligence for Reverse Referral Between A Hospital Emergency Department and A Primary Urgent Care Center,” Front. Digit. Heal., vol. 7, no.

March, pp. 5–10, 2025, doi: 10.3389/fdgth.2025.1546467.

A. I. Taloba, R. M. A. El-aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, “Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning,” J. Healthc. Eng., vol. 2022, no. 1, pp. 1–10, 2022, doi: <https://doi.org/10.1155/2022/7969220> Research.

S. Melhem, A. Al-Aiad, and M. S. Al-Ayyad, “Patient Care Classification Using Machine Learning Techniques,” in International Conference on Information and Communication Systems (ICICS), 2021, pp. 1–6. doi: 10.1109/ICICS52457.2021.9464582.

F. Chollet, Deep Learning with Python. United States: Manning Publications Co, 2018. doi: 10.4018/978-1-7998-3095-5.ch003.

A. Masbakhah, U. Sa’adah, and M. Muslikh, “Heart Disease Classification Using Random Forest and Fox Algorithm as Hyperparameter Tuning,” J. Electron. Electromed. Eng. Med. Informatics, vol. 7, no. 4, pp. 964–976, 2025, doi: 10.35882/jeeemi.v7i4.932.

M. Awad and R. Khanna, Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, no. April 2015. 2015. doi: 10.1007/978-1-4302-5990-9.

H. Huang, R. Jia, J. Liang, J. Dang, and Z. Wang, “Wind Power Deterministic Prediction and Uncertainty Quantification Based on Interval Estimation,” J. Sol. Energy Eng. Trans. ASME, vol. 143, no. 6, pp. 1–12, 2021, doi: 10.1115/1.4051430.

L. Costa et al., “Multilayer Perceptron,” in Introduction to Computational Intelligence, no. January, IEEE, 2023, pp. 1–6. doi: 10.1145/3090354.3090427.

H. Yin, “Enhancing Directional Accuracy in Stock Closing Price Value Prediction Using a Direction-Integrated MSE Loss Function,” in Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023), 2023, pp. 119–126. doi: 10.5220/0012810200003885.

T. O. Hodson, “Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Tot,” *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.

A. Danujan, T. Nithusikan, Athauda, Hearth, and Senanayake, “Predicting Inpatient Bed Demand Using Machine Learning,” in *Young Members’ Technical Conference 2024*, 2023, pp. 150–157. [Online]. Available: [https://www.researchgate.net/publication/391584272\\_Predicting\\_Inpatient\\_Bed\\_Demand\\_Using\\_Machine\\_Learning](https://www.researchgate.net/publication/391584272_Predicting_Inpatient_Bed_Demand_Using_Machine_Learning)

B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, “Predicting Health Care Costs Using Evidence Regression,” in *MDPI*, 2019, pp. 1–13. doi: 10.3390/proceedings2019031074.

K. Alghatani, N. Ammar, A. Rezgui, and A. Shaban-nejad, “Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation,” *JMIR Med. INFORMATICS*, vol. 9, no. 5, pp. 1–23, 2021, doi: 10.2196/21347.

J. An, M. Jung, S. Ryu, Y. Choi, and J. Kim, “Analysis of Length of Stay for Patients Admitted to Korean Hospitals Based on The Korean National Health Insurance Service Database,” *Informatics Med. Unlocked*, vol. 37, no. October 2022, p. 101178, 2023, doi: 10.1016/j.imu.2023.101178.

D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon, “Predicting Inpatient Flow at A Major Hospital using Interpretable Analytics,” *Manuf. Serv. Oper. Manag.*, vol. 2020, no. 1, pp. 1–42, 2020, doi: <https://doi.org/10.1101/2020.05.12.20098848>.

### Most read articles by the same author(s)

- Suryani, Angger Totik Prasetyo, Gandung Triyono, [COMPARISON OF SAW AND TOPSIS METHODS TO DETERMINE THE BEST SERVICE DESK AGENT](#), *Jurnal Teknik Informatika (Jutif)*: Vol. 5 No. 1 (2024): JUTIF Volume 5, Number 1, February 2024
- Muhamad Fadel, Kanasfi, Zainal Arifin, Gandung Triyono, [APPLICATION OF ENSEMBLE METHOD FOR EMPLOYEE](#)

[TURNOVER PREDICTIONS IN FINANCIAL SERVICES COMPANY](#) , [Jurnal Teknik Informatika \(Jutif\): Vol. 5 No. 3 \(2024\): JUTIF Volume 5, Number 3, June 2024](#)

- Greghar Juan Tjether Maruanaya, Gandung Triyono, Rita Fransina Maruanaya, [DEVELOPMENT OF A STOCK PURCHASE RECOMMENDATION SYSTEM APPLICATION](#) , [Jurnal Teknik Informatika \(Jutif\): Vol. 5 No. 4 \(2024\): JUTIF Volume 5, Number 4, August 2024 - SENIKO](#)
- Ajar Parama Adhi, Khairil Umuri, Gandung Triyono, [SENTIMENT ANALYSIS AND ENTITY DETECTION ON NEWS HEADLINES TO SUPPORT INVESTMENT DECISIONS](#) , [Jurnal Teknik Informatika \(Jutif\): Vol. 5 No. 6 \(2024\): JUTIF Volume 5, Number 6, Desember 2024](#)

Copyright © Jutif Unsoed 2024

[Jutif's Stats](#) [Hit counters](#)

Indexed By :



Jurnal Teknik Informatika (JUTIF)

P-ISSN : 2723-3863 | e-ISSN : 2723-3871

This work is licensed under a

