JUTIF
Jurnal Teknik Informatika

P-ISSN: 2723-3863
E-ISSN: 2723-3871

VOLUME 7 | NUMBER 3 | JUNE 2026

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

JUTIF
Jurnal Teknik Informatika
p-ISSN : 2723-3863
e-ISSN : 2723-3871



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


Jurnal Teknik Informatika (JUTIF) UNSOED Volume 7, Number 3, June 2026 was published on June 15, 2026. JUTIF in this edition has received quite a lot of article submissions, but in the process some of the best articles have been selected according to the results of the review. TThis edition published article from several affiliations, including :

**Universiti Tun Hussein Onn Malaysia (UTHM) (Malaysia), University of Technology Malaysia (Malaysia), Albukhary International University (Malaysia), King Fahd University of Petroleum and Minerals (Saudi Arabia), Nnamdi**

MAKE A SUBMISSION

Download Template

Azikiwe University (Nigeria), University of Dili (Undil) Timor Leste (Timor-Leste), National Taiwan University of Science and Technology (Taiwan), University of Illinois (United States), Monash University (Australia), Universitas Sains dan Teknologi Indonesia (Indonesia), Universitas Islam Negeri Sunan Kalijaga (Indonesia), STIKOM Tunas Bangsa (Indonesia), Universitas Nusa Nipa (Indonesia), Universitas Dhyana Pura (Indonesia), Institute of Technology and Health (Indonesia), University of Sugeng Hartono (Indonesia), Universitas Ciputra (Indonesia), Sekolah Tinggi Ilmu Komputer PGRI Banyuwangi (Indonesia), Universitas Amikom Yogyakarta (Indonesia), Universitas Cokroaminoto Palopo (Indonesia), University of Pembangunan Nasional Veteran (Indonesia), Universitas Harkat Negeri (Indonesia), Universitas Diponegoro (Indonesia), Universitas Sebelas Maret (Indonesia), Universitas Islam Lamongan (Indonesia), Institut Teknologi Bacharuddin Jusuf Habibie (Indonesia), Universitas Handayani Makassar (Indonesia), Universitas Andalas (Indonesia), Lambung Mangkurat University (Indonesia), Budi Luhur University (Indonesia), Telkom University (Indonesia), Institut Teknologi Sepuluh Nopember (Indonesia), Dinamika Bangsa University (Indonesia), Universitas Jenderal Soedirman (Indonesia), Universitas Negeri Makassar (Indonesia), Politeknik Negeri Cilacap (Indonesia), Politeknik Negeri Semarang (Indonesia), SEAMEO SEAMOLEC (Indonesia), Universitas Cendana (Indonesia), Universitas Muhammadiyah Yogyakarta (Indonesia), Universitas Syiah Kuala (Indonesia), STMIK Kaputama (Indonesia), Universitas Dian Nuswantoro (Indonesia), Institut Bisnis dan Teknologi Kalimantan (Indonesia), Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan (Indonesia), STMIK IKMI Cirebon (Indonesia), Universitas Muhammadiyah Bima (Indonesia), Institut Teknologi Garut (Indonesia), Universitas Pendidikan Indonesia (Indonesia), Airlangga University (Indonesia), Utpadaka Swastika University (Indonesia), Mulawarman University (Indonesia), Universitas Muhammadiyah Bengkulu (Indonesia),



</a

**..:: ISSN ::..**

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

**..:: ABOUT JUTIF ::..**

Focus and Scope

Editorial Team

Reviewers

Author Guidelines

Peer Review Process

**Universitas Sulawesi Barat (Indonesia), Universitas Islam Madura (Indonesia), Ahmad Dahlan University (Indonesia), Universitas Muhammadiyah Malang (Indonesia), Universitas Yapis Papua (Indonesia), Adisutjipto Institute of Aerospace Technology (Indonesia), Universitas Muhammadiyah Surakarta (Indonesia), University of Trunojoyo Madura (Indonesia), Universitas Potensi Utama (Indonesia), Universitas Putra Indonesia YPTK Padang (Indonesia), Putra Indonesia University (Indonesia), Esa Unggul University (Indonesia), Universitas PGRI Semarang (Indonesia), Bina Nusantara University (Indonesia), Politeknik Negeri Medan (Indonesia), dan Universitas Negeri Semarang (Indonesia).**

**PUBLISHED:** 2026-06-15

## FULL ISSUE

COVER JUTIF VOL 7 NO 3

PREFACE AND TABLE OF CONTENTS

## ARTICLES



Open Access Policy

Licensing & Copyright

Archiving

Plagiarism Policy

Publication Frequency

Publication Ethics

Article Processing Charge

Generative AI Policy

Indexing

Journal's History

Contact

## Predicting Mental Health Status using a Fine-Tuned CNN-LSTM Hybrid Model

Agustin, Junadhi, Susi Erlinda, Triyani Arita Fitri, Lusiana Efrizoni 2978-2993

DOWNLOAD PDF

## Hybrid Cryptography-Steganography Scheme Based on Camellia-256 and LSB for Enhanced Security and Imperceptibility of Secret Messages

Imam Prayogo Pujiono, Eko Hari Rachmawanto, Christy Atika Sari, Said Fachri Ariza, Isnaeni Kholifatun 2491-2505

DOWNLOAD PDF

## Forecasting Nutrient Concentration Dynamics in Hydroponic Lettuce Cultivation Using a Hybrid Fuzzy Time Series and Long Short-Term Memory Approach for Internet of Things–Based Systems

Muh. Agus, Alvian Tri Putra Darti Akhsa, Ilham Ali Marka M, Muhammad Fadel Hasyim 2277-2293

DOWNLOAD PDF

## Comparative Analysis of Hyperparameter Optimization Methods for LSTM in Cryptocurrency Price Prediction: An Application to TRX–USD

Dasril Aldo, Muhammad Raafi'u Firmansyah, Muhammad Afrizal Amrustian 2340-2349

DOWNLOAD PDF

## Optimization of ShuffleNetV2 Using Self-Knowledge Distillation for Cocoa Fruit Disease Classification



## INFORMATION

DOWNLOAD PDF

## Enhancing Flood Area Segmentation in Remote Sensing Images Using Hybrid Attention Mechanism on DeepLabV3+ with ResNet-50 Backbone

Annisa Syifaul Ummah, Esti Suryani, Herdito Ibnu Dewangkoro 2246-2258

DOWNLOAD PDF

## Implementation of Moving Average and Weighted Moving Average for Forecasting Palm Oil Harvest and Income in a Web-Based GIS System

Elvia Andriyani, Bambang Agus Herlambang, Khoiriya Lathifa 2891-2905

DOWNLOAD PDF

## Model-Driven Engineering for ARrupiah Cultural AR App: Kano Model and Qualitative User Experience Evaluation

Gerson Feoh, I Made Dwi Ardiada, Gabriel Firsta Adnyana, I Gede Hendrayana 2061-2078

DOWNLOAD PDF

## Optimizing Naive Bayes for Sentiment Analysis of M-Passport Reviews Using N-Gram and Synthetic Minority Over-sampling Technique

Devia Kartika, Sarjon Defit 2798-2811

DOWNLOAD PDF

## Optimization of Machine Learning Model using Grid and Random Search Algorithms for Predicting Student Dropout

M. Faris Al Hakim, Siti Wahyuni, Kholiq Budiman, Aditya Marianti, Bambang Eko Susilo, Nuni Widiarti, Sri Sukaesih, Rifaatunnisa

3012-3024

DOWNLOAD PDF

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

Gandung Triyono, Billy Marentek, Mohammad Syafrullah

2323-2339

DOWNLOAD PDF

## Development of a Hybrid Machine Learning-Based E-Commerce Chatbot Using Jaccard Similarity and K-Nearest Neighbor for Accurate Intent Classification

Andrian Sah, Andi Ilham, Rasna, Siti Nurhayati

2720-2733

DOWNLOAD PDF

## VGG-16 Transfer Learning for Accurate Classification of Three Local Durian Varieties Using Leaf Morphology Images

Ahmad Haikal Nuqqy Zahhar, I Gede Susrama Mas Diyasa, Made Hanindya Prami Swari

2165-2188

DOWNLOAD PDF

## Prediction of Indonesian Banking Stock Prices Using a Hybrid LSTM and XGBoost Model with Optuna Based Hyperparameter Optimization

DOWNLOAD PDF

## Comparative Evaluation of Linear Regression and Ensemble Learning Models for Daily Calorie Prediction Using a Public Lifestyle Dataset with Structured Preprocessing and Recursive Feature Elimination

Yunandra Wahyu Utama, Majid Rahardi 2841-2856

DOWNLOAD PDF

## Constructing a Part-of-Speech Tagging based on Lexicon and Rule-based for Sundanese Corpus

Ade Sutedi, Ayu Latifah, Novan Rodiansyah, Yayat Sudaryat 2534-2543

DOWNLOAD PDF

## Sentiment Analysis Of E-Commerce Reviews Using Fine-Tuned Indobert With Class Weights Strategy

Abdan Syakura, Dewi Soyusiawaty 2690-2703

DOWNLOAD PDF

## Classification of Watermelon Flavor Using Artificial Neural Network with Color, Texture, and Shape Features

Muh Faqih S Musgamy, Muhammad Risaldi, Ayu Safitri, Andi Baso Kaswar, Muhammad Fajar B, Jumadi M Parenreng 2544-2560

DOWNLOAD PDF

DOWNLOAD PDF

## Software Development Cost Prediction Using XGBoost with Optuna-Based Hyperparameter Optimization on the COCOMO Dataset

Eddy Maryanto, Bangun Wijayanto, Swahesti Puspita Rahayu, Dwi Kurnia Wibowo 3064-3080

DOWNLOAD PDF

## Image Cryptography Process Using Arnold's Cat Map And Henon Map Algorithms

Moch Azhar Al Ghifari, Bayu Surarso, Aris Sugiharto 2225-2245

DOWNLOAD PDF

## Banana Leaf Disease Classification Using CNN Feature Extraction and Naive Bayes Algorithm

Moh. Badri Tamam, Januario Freitas Araujo , Anwari 2646-2659

DOWNLOAD PDF

## Comparative Analysis of the Performance of Random Forest and CatBoost for Air Quality Prediction Based on Meteorological Factor

Nirsal, Nurchaerani Kadir 2154-2164

DOWNLOAD PDF

## Artificial Intelligence-Based Aircraft Detection for Enhanced Aviation Safety and Air Traffic Management



## Development of Smart Study Web Application for Classifying Student Material Understanding Levels Using Naive Bayes Classifier



## Optimizing Breast Cancer Classification: SVM and Random Forest with Hybrid Hyperparameter Tuning and Feature Selection



## Hybrid Unsupervised-Supervised Learning for Housing Submarket Segmentation and Price Prediction in Surabaya Urban Areas



## Information Gain-Based Feature Selection and Machine Learning Classification

DOWNLOAD PDF

## Comparative Analysis of Temporal Fusion Transformer and Long Short-Term Memory Architecture Resilience in Predicting Solana Price Volatility Across Different Market Phases

Mahdy Eka Putra, Tanty Oktavia 2935-2943

DOWNLOAD PDF

## Detection of Coffee Leaf Diseases Using Deep Learning to Support Digitalization and Smart Agriculture

Siti Mutmainah, Zumhur Alamin 2517-2533

DOWNLOAD PDF

## Performance Comparison Of K-Nearest Neighbors And Decision Tree Algorithms With Random Oversampling For Imbalanced Heart Disease Classification

Dita Yustianisa, Farid Wajidi, Wawan Firgiawan, Adinda Gama Sholeha 2633-2645

DOWNLOAD PDF

## Certainty Factor Algorithm Approach for Early Stage of Cattle Disease Diagnosis Using Mobile-Based Expert System

Budy Satria, Nurfiah, Bima Prakasa 2294-2306

DOWNLOAD PDF

**Jurnal Teknik Informatika (JUTIF)**

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

**Jurnal Teknik Informatika (JUTIF)**
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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

**Gandung Triyono*[1], Billy Marentek[2], Mohammad Syafrullah[3]**

[1]Information System, Faculty of Information Technology, Budi Luhur University, Indonesia
[2,3]Master of Computer Science, Faculty of Information Technology, Budi Luhur University, Indonesia

Email: [1]gandung.triyono@budiluhur.ac.id

Received : Oct 30, 2025; Revised : Dec 15, 2025; Accepted : Jan 19, 2026; Published : Jun 15, 2026

**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*



## 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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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 ≈ 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 ≥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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

~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.

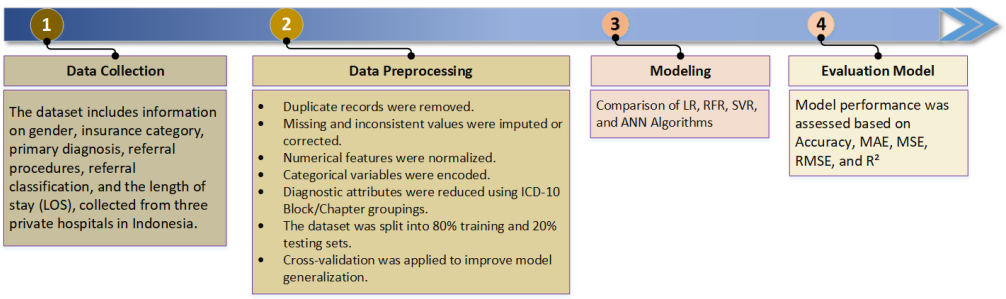


Figure 1. Workflow of LOS Prediction Model Development

### 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.

#### 2.3.1. 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):

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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (3)$$

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):

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

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

The modeling process using the linear regression algorithm begins by determining the input variables ($X_1 \ldots X_n$), 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 $X_1 \ldots X_n$ and the outcome Y. During this process, the coefficients $\beta_0, \beta_1 \ldots \beta_n$ are calculated by minimizing the residual error ε, enabling the model to generate a regression line that best approximates the actual length of stay experienced by newborn patients.

### 2.3.2. 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].

### 2.3.3. 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-

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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· ·· N}, the wind power prediction of SVR method can be given as y = f(x) =ωφ(x) + bx, where w is the weight, b is the parameter of bias, and φ(x) represents the high- dimensional space. To obtain parameters w and b, the prediction of SVR model with structure risk minimization principle can be trans- formed to a mathematical optimization problem and solved using Lagrange multiplier [23]. The optimization problem is presented as follows (6):

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

Here, $\boldsymbol{Y_i}$and $\boldsymbol{Y_i^*}$are the Lagrange multipliers, and $\boldsymbol{K(X, X_i)}$ 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.

### 2.3.4. Artificial Naural 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 $\boldsymbol{j}$ is calculated as (7):

$$\boldsymbol{Z_{inj} = V_{0j} + \sum_{i=1}^{n} X_i V_{ij}\, X_i V_{ij}} \qquad (7)$$

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

$$\boldsymbol{Z_j = f(Z_{inj}), f(x) = \frac{1}{1+e^{-x}}} \qquad (8)$$

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

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

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

The output $\boldsymbol{Y_k}$is then compared with the target value $\boldsymbol{T_k}$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):

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

with the sigmoid derivative (12):

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

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

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

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

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

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

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

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

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

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

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

3) Testing Phase

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

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

## 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):

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

where n represents the sample size 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[18], [26], as expressed in the following (21):

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

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):

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

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):

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

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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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).

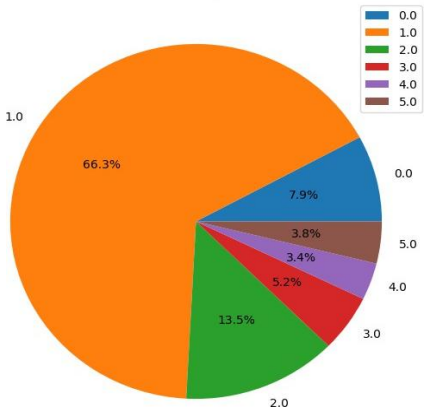


Figure 2. Distribution of LOS Data

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.

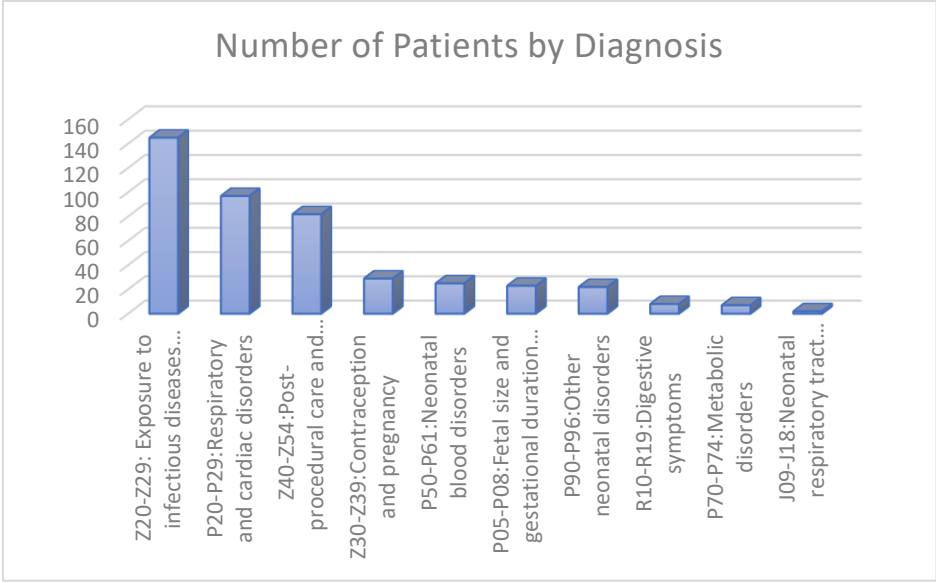


Figure 3. Data Distribution for Diagnosis

### 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.

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

- 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.

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 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.

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

In Table 3, the hyperparameters Kermel, 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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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

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 ≥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 |

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 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," *Apllied 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:

Vol. 7, No. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

10.20944/preprints202302.0315.v1.

[5] K. Stone, R. Zwiggelaar, 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.

[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," *Apllied 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. 3, June 2026, Page. 2323-2339
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5410

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 ofthe 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

[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.