



AI and Data Science

Technology, Innovation & Use Cases in Indonesia

Editor: RUDI RUSDIAH

Opening Remark: BAMBANG BRODJONEGORO

JENDERAL (PURN) TNI MOELDOKO | ANANG ACHMAD LATIF

Authors: HAMMAM RIZA | WENDI USINO | ON LEE

INDRA UTOYO | JAYAPRAWIRYA DIAH | FRADO SIBARANI | ROHIT KUMAR

AI and Data Science

Technology, Innovation & Use Cases in Indonesia



Editor: RUDI RUSDAH

**Opening Remark: BAMBANG BRODJONEGORO
JENDERAL (PURN) TNI MOELDOKO | ANANG ACHMAD LATIF**

**Authors: HAMMAM RIZA | WENDI USINO | ON LEE
INDRA UTOYO | JAYAPRAWIRYA DIAH | FRADO SIBARANI | ROHIT KUMAR**

Penerbit: Perkumpulan Basis Data Indonesia



AI AND DATA SCIENCE **Technology, Innovation & Use Cases in Indonesia**

Editor: Dr. Rudi Rusdiah

ISBN: 978-623-92672-2-3

Desain sampul: Samtoryke

Layout: Samuel R. & Komite.ID Magazine Team

© **Penerbit:** Perkumpulan Basis Data Indonesia

Redaksi:

Asosiasi Big Data & AI Indonesia (ABDI)

Golden Plaza A38, Jl. RS. Fatmawati No. 15

Jakarta Selatan – Indonesia, 12410

Ph. 021 – 75900091/93

Email: abdi@indopc.co.id

www.abdi.id

Cetakan pertama, November 2020

Hak cipta dilindungi oleh undang-undang.

Dilarang memperbanyak atau mengutip sebagian atau seluruh isi buku tanpa seizin tertulis dari Penerbit.

Dicetak oleh Percetakan PT. Gramedia Jakarta
isi di luar tanggung jawab percetakan.

Contents

KATA SAMBUTAN

MENTERI RISET DAN TEKNOLOGI/KEPALA BRIN - PROF. BAMBANG PS. BRODJONEGORO, PH.D.	xi
KEPALA STAF KEPRESIDENAN REPUBLIK INDONESIA - JENDERAL (PURN) TNI DR. H. MOELDOKO	xii
KEPALA BADAN PENGKAJIAN DAN PENERAPAN TEKNOLOGI - DR. IR. HAMMAM RIZA, M.SC, IPU, AER.....	xiv

KATA PENGANTAR

DIRUT BADAN AKSESIBILITAS TELEKOMUNIKASI DAN INFORMASI (BAKTI) - IR. ANANG ACHMAD LATIF, M.SC.....	xvii
--	------

PRA KATA.....	XX
---------------	----

CHAPTER 1: PENGENALAN TENTANG ARTIFICIAL INTELLIGENCE.....	1
--	---

MEMAHAMI ARTIFICIAL INTELLIGENCE (AI)	3
---	---

By. Rudi Rusdiah - Micronics Internusa

PENDAHULUAN	3
TABEL MEMAHAMI BERBAGAI JENIS INTELIJEN	4
MENEMUKAN 4 CARA DEFINISI AI	5
GENERAL (UMUM) & WEAK (LEMAH) AI.....	7
SEJARAH DARI AI.....	8
DATA SCIENCE & ENGINEER.....	10
ALGORITHMMA MERUBAH PERMAINAN DAN POLITIK DARI DATA.....	11
HARDWARE (H/W) & KEKUATAN KOMPUTASI	12
TATAKELOLA AI (ARTIFICIAL INTELLIGENT) & LOWONGAN TENAGA KERJA	14

STRATEGI NASIONAL KECERDASAN ARTIFISIAL UNTUK MELAKUKAN LOMPATAN INOVASI TEKNOLOGI MENUJU VISI INDONESIA 2045	15
---	----

By. Hammam Riza dan Hary Budiarto - Badan Pengkajian dan Penerapan Teknologi (BPPT)

PENDAHULUAN	15
TEKNOLOGI KECERDASAN ARTIFISIAL	15
PENYIAPAN LOMPATAN INOVASI TEKNOLOGI KECERDASAN ARTIFISIAL.....	19
INFRASTRUKTUR KECERDASAN ARTIFISIAL DI BPPT.....	20
KECERDASAN ARTIFISIAL UNTUK MESIN PENERJEMAH	21
KECERDASAN ARTIFISIAL UNTUK PENGENALAN WICARA (SPEECH RECOGNITION)	23
KECERDASAN ARTIFISIAL UNTUK SISTEM IDENTIFIKASI BERBASIS MULTIMODAL BIOMETRIK	25
KECERDASAN ARTIFISIAL (KA) UNTUK ALAT DIAGNOSIS MALARIA DARI CITRA MIKROSKOPIS APUSAN DARAH.....	26
KECERDASAN ARTIFISIAL (KA) UNTUK MITIGASI BENCANA BANJIR	27

KECERDASAN ARTIFISIAL (KA) UNTUK PENANGGULANGAN REDUKSI BENCANA (PRB) KEBAKARAN HUTAN DAN LAHAN (KARHUTLA)	29
INOVASI TEKNOLOGI KECERDASAN ARTIFISIAL (KA) UNTUK COVID-19	31
STRATEGI NASIONAL KECERDASAN ARTIFISIAL	32
REFERENSI	37
CHAPTER 2: FOCUS PADA TEKNOLOGI DASAR ARTIFICIAL INTELLIGENCE	39
DIGITAL RETRANSFORMATION IN INDONESIA	41
<i>By. OnLee - GDP Venture, GDP Labs and Kaskus</i>	
INTRODUCTION	41
DIGITAL RETRANSFORMATION USING AI	42
AI OPPORTUNITIES AND CHALLENGES	42
OPPORTUNITIES	42
CHALLENGES	43
TECHNOLOGY-DRIVEN COMPANIES	44
AI LANDSCAPE	47
GDP VENTURE INVESTMENTS IN AI	47
BUSINESS USE CASE SCENARIOS	48
EKYC	48
REGULAR TECHNOLOGY (REGTECH)	48
WORK FROM HOME & WORK FROM OFFICE	49
CONCLUSION	50
PABRIK AI: OTOMATISASI PROSES MANAJEMEN OLEH AI	51
<i>By. Rudi Rusdiah - KOMITE.ID</i>	
PENDAHULUAN	51
PERANAN EVOLUSI IT, DT DAN AI	51
INTI OTAK (CORE) DARI PABRIK AI	51
CHAT & VOICE BOT (ROBOT) AI	52
HUMANOID BOTS NOW?	52
STRONG & WEAK AI	53
AI WARFARE (THE FIFTH DOMAIN): DRONE & STRONG AI	53
AI FACTORY	54
TANTANGAN MEMBANGUN AI FACTORY DARI BISNIS TRADISIONAL	55
MELETAKKAN AI DI INTI DARI BISNIS ENTERPRISE	56
ARSITEKTUR YANG KONSISTEN & JELAS	56
STUDI KASUS: TRANSFORMASI AI DARI MICROSOFT	57
STUDI KASUS: AI BORN ANT FINANCE, UPAYA STARTUP MENANG BERKOMPETISI DENGAN BISNIS LEGACY?	57
PERJALANAN DENGAN R DARI PENGETAHUAN DASAR DATA SCIENCE MENUJU MACHINE LEARNING (ML)	61
<i>By. Rudi Rusdiah - Asosiasi Big Data & AI (ABDI)</i>	
PENDAHULUAN	61
MENGAPA R DAN R STUDIO?	61

DATA VISUALIZATION (VISUALISASI DATA)	62
STATISTIK DENGAN R	63
PROBABILITAS (PROBABILITY)	64
DATA SOURCE (SUMBER)	65
DATA WRANGLING (CLEANSING)	66
WEB SCRAPING (WS)/WEB HARVESTING	67
STRING PROCESSING	68
TEXT MINING	68
SENTIMEN ANALYSIS (SA).....	68
BEKERJA DENGAN ALGORITMA	69
MEMAHAMI APA ARTI ALGORITMA.....	70
MENGAPA AI MEMBUHTUKAN MACHINE LEARNING?	70
NOTATION.....	71
STUDI KASUS ML KODE POS: MENJELASKAN PREDIKTOR, HASIL DAN FITUR	73
MEMBUAT DAN EVALUASI ALGORITMA.....	75
PENGENALAN MENGENAI R UNTUK LATIHAN EVALUASI ALGORITMA	75
KASUS MENGEVALUASI ALGORITMA JENIS KELAMIN DARI DATA TINGGI BADAN	75
CHAPTER 3: USE CASE & SOLUSI AI BUSINESS DOMAIN	81
BRI-BRAIN, ARTIFICIAL INTELLIGENCE TERPUSAT MILIK BANK BRI	83
<i>By. Indra Utoyo dan Tio Anta Wibawa - Bank Rakyat Indonesia, Tbk.</i>	
PENDAHULUAN	83
AI DAPAT MEMBANTU PERCEPATAN PERTUMBUHAN BISNIS PERBANKAN	85
INOVASI BIG DATA HARUS SELALU BERLANJUT	86
BRI BRAIN, OTAK PENGAMBILAN KEPUTUSAN BANK BRI	87
FRAMEWORK BRIBRAIN	91
DAMPAK BRI BRAIN TERHADAP BISNIS BANK BRI	92
SUSTAINABLE DEVELOPMENT BRI-BRAIN.....	94
REFERENSI	95
WHAT AND HOW WE IMPLEMENT ARTIFICIAL INTELLIGENCE IN BCA.....	97
<i>By. Hendra Lembong, Adhitya B. Tirtohadiguno and Jayaprawirya Diah - Bank Central Asia, Tbk.</i>	
INTRODUCTION	97
THE BEGINNING, AND BEYOND	97
COMPUTER VISION.....	99
NATURAL LANGUAGE PROCESSING (NLP) AND VOICE BIOMETRICS	99
CUSTOMER PROFILING AND MARKETING.....	100
FRAUD AND ANOMALY DETECTION.....	100
AI IMPLEMENTATION KEY SUCCESS FACTORS IN BCA.....	100
PEOPLE : TALENT, ORGANIZATION, AND CULTURE	100
PROCESS: UNDERSTANDING, ITERATION, DATA, AND DEPLOYMENT	103
TECHNOLOGY: THE RIGHT SOLUTION FOR THE RIGHT PROBLEM	104
DATA : QUANTITY, QUALITY, AND INTEGRATION	105

SUMMARY AND PERSPECTIVES FOR THE FUTURE	105
COVID-19 EPIDEMIC ANALYSIS USING MACHINE LEARNING & DEEP LEARNING ALGORITHMS	107
<i>By. Sonali Agarwal, Narinder S. Punn, Sanjay K. Sonbhadra and M. Syafrullah - IIT Allahabad and UBL</i>	
ABSTRACT	107
INTRODUCTION	107
COVID-19 TRANSMISSION STAGES	109
RELATED WORK.....	111
DATASET DESCRIPTION.....	112
EPIDEMIC ANALYSIS	113
TRAINING AND TESTING	113
RESULTS AND DISCUSSION.....	114
CONCLUSION.....	115
REFERENCES.....	116
MENGGUNAKAN ARTIFICIAL INTELLIGENCE DI BIDANG HEALTHCARE.....	117
<i>By. Frado Sibarani - Cognixy.ai</i>	
PENDAHULUAN	117
CONTOH KASUS: MENDETEKSI COVID-19 MELALUI X-RAY IMAGES DENGAN MENGGUNAKAN AI.....	120
SETUP.....	121
BUILD DATASET	121
COVID XRAY DATASET.....	121
BUILD NORMAL XRAY DATASET	122
PLOT X-RAYS.....	122
DATA PREPROCESSING	124
MODEL	124
TRAINING	125
PLOT TRAINING METRICS	126
EVALUATION.....	126
CONFUSION MATRIX.....	127
CHAPTER 4: USE CASE TEKNOLOGI AI BUSINESS DOMAIN.....	129
UCOACH: BEGIN WITH THE END IN MIND	131
<i>By. Frans Budi Pranata dan Gendro Salim - Yupi Gummy Candy dan PT. UCOACH Djivasrana Grahasada</i>	
PENDAHULUAN	131
BIG DATA.....	133
MINDMAP	134
LATERAL THINKING	134
DATA MAPPING	135
DATA ANALYTIC	136
ARTIFICIAL INTELLIGENCE (AI).....	137
PHYTON & BLOCKCHAIN.....	138

AI & DATA SCIENCE	138
APLIKASI ARTIFICIAL INTELLIGENCE (AI) DI KOMUNITAS MASYARAKAT	141
<i>By. M. Windy Trihantoro, Khoirunnisa, Amin Muzaeni dan Yogi A. Subekhti - Mahasiswa S2 UBL</i>	
PENDAHULUAN	141
PENGUNAAN AI DALAM KOMPUTER APLIKASI	142
PENGENALAN TIPE APLIKASI AI SECARA UMUM	142
MELIHAT BAGAIMANA AI MEMBUAT APLIKASI YANG RAMAH	144
AI MEMBERIKAN SARAN SECARA OTOMATIS.....	145
PENGUNAAN AI DALAM PROSES AUTOMASI	146
SOLUSI AI UNTUK SESUATU YANG MEMBOSANKAN	146
TANTANGAN AI UNTUK MENJAGA KEAMANAN	150
ALGORITMA DAN APLIKASI MACHINE LEARNING	155
<i>By. Irwan Susanto, Ummu Habibah R. dan Suwardiman - Mahasiswa S2 UBL</i>	
PENDAHULUAN	155
APLIKASI MACHINE LEARNING	156
KONTROVERSI MACHINE LEARNING	156
MACHINE LEARNING PROCESS.....	157
TAHAPAN PEMBANGUNAN MACHINE LEARNING	157
KOMPONEN KUNCI PADA ALGORITMA MACHINE LEARNING.....	158
MENINGKATKAN PERFORMA MACHINE LEARNING	158
METODE PEMBELAJARAN	161
METODE PEMBELAJARAN MACHINE LEARNING	161
SUPERVISED LEARNING	161
UNSUPERVISED LEARNING.....	162
REINFORCEMENT LEARNING.....	163
PENERAPAN MACHINE LEARNING.....	164
PENERAPAN PENDEKATAN SUPERVISED LEARNING	164
KLASIFIKASI DAN REGRESI.....	165
KNN CLASIFFIER.....	167
PENUTUP	173
STATE OF THE ART PERSPECTIVE ON AI APPLICATION IN INDONESIA	175
<i>By. Rizaldi Sistiabudi and Angga Wirapraditya Rhamdani - PT Cakra Tekno Nusantara (CAKRA.ai)</i>	
INTRODUCTION	175
AI APPLICATIONS IN INDONESIA.....	176
PUBLIC SECTORS	176
BANKING & FINANCIAL SERVICE INDUSTRY	179
AGRICULTURE	182
BUILDING BLOCKS OF CAKRA TEKNO NUSANTARA'S AI PLATFORM	183
CONCLUSION.....	185

AIMYPET®: CASE STUDY OF AN AI ROBOTIC VENTURE	187
<i>By. Peng Chan and Will Smith - Aimy Robotics & Global Management Group</i>	
ABSTRACT	187
THE AI REVOLUTION	188
SAY HELLO TO AIMY®.....	189
USES & APPLICATIONS FOR ROBOTIC DOG	190
AS A PERSONAL TRUSTED COMPANION	190
AS A TRUSTED HEALTHCARE GIVER	190
MARKET DEMAND AND ACCEPTANCE FOR ROBOTIC DOGS.....	191
IS THE WORLD READY FOR AIMY?	193
TECHNOLOGY FOR THE 4-LEGGED EDGE COMPUTING SENSOR PLATFORM	193
TECHNOLOGY PARTNERS	194
ALL-TERRAIN 4-LEGGED PLATFORM	194
NVIDIA ROBOT JETSON SYSTEMS	194
MAPPING SURROUNDINGS USING 3D IMAGING	195
EMOTION CHIP	195
TECHNOLOGY INTEGRATION	195
CONCLUSION.....	196
PEMANFAATAN TEKNOLOGI AI- NATURAL LANGUAGE PROCESSING (NLP) UNTUK BISNIS	197
<i>By. Ayu Purwarianti - Prosa.ai & AI Center ITB, PUI-PT AI-VLB</i>	
PENGERTIAN	197
MANFAAT TEKNOLOGI NLP (PEMROSESAN TEKS)	198
PENGUNAAN TEKNOLOGI NLP DI BISNIS	199
REINFORCEMENT LEARNING: SOLUSI KECERDASAN BUATAN	205
<i>By. Anton Wardaya, Kevin Daniel Pantasdo, Christopher Matthew & Prawira S. Darma - Wardaya College</i>	
PENDAHULUAN	205
ELON MUSK TAKUT AI MENGAMBIL ALIH DUNIA	205
REINFORCEMENT LEARNING - SOLUSI KECERDASAN BUATAN	206
CONTOH-CONTOH REINFORCEMENT LEARNING	207
RL UNTUK PENGIRIMAN LOGISTIK DARI PABRIK	207
RL UNTUK KEPUTUSAN INVESTASI KEUANGAN	207
RL UNTUK BIDANG MEDIS.....	208
RL UNTUK SEGALA MASALAH	208
TEORI DASAR REINFORCEMENT LEARNING	208
1. POLICY-BASED APPROACH	211
2. VALUE-BASED APPROACH.....	212
TANTANGAN YANG DIHADAPI REINFORCEMENT LEARNING	212
SEJARAH DAN PERKEMBANGAN REINFORCEMENT LEARNING	213
NEURAL NETWORKS DALAM UNTUK PENGENALAN SUARA	216
NEURAL NETWORKS GOOGLE MENGENALI VIDEO KUCING	216
KEDATANGAN ALEXNET	217

CNN MENCAIPI TINGKAT KESALAHAN TERENDAH SEPANJANG MASA	217
KONKLUSI	218
REFERENSI	219
TARGETING CLASS IMBALANCE PROBLEM IN CREDIT CARD FRAUD DETECTION USING GENERATIVE ADVERSARIAL NETWORK SYNTHETIC OVERSAMPLING FRAMEWORK	221
<i>By. Sonali Agarwal, Narinder S. Pun, Sanjay K. Sonbhadra & Wendi Usino - IIIT Allahabad and UBL</i>	
ABSTRACT	221
INTRODUCTION	222
IMBALANCED LEARNING	222
OVERVIEW	223
RELATED WORK.....	224
PROPOSED ARCHITECTURE AND METHODOLOGY	225
SYNTHETIC DATA GENERATION USING GANSOF	226
CREDIT CARD FRAUD DETECTION.....	228
EXPERIMENTATION AND RESULTS	228
DATASET	228
TRAINING AND EVALUATION.....	228
RESULTS AND DISCUSSION.....	230
CONCLUSION.....	230
REFERENCES.....	231
MENGENAL POTENSIA: MARKETPLACE KOMODITAS ALAM PERTAMA DI INDONESIA.....	233
<i>By. Febi M Faizal, Qori Utama dan Franz Budi Pratama - Pontesiana.com</i>	
PENDAHULUAN	233
ARTIFICIAL INTELLIGENCE DALAM MARKETPLACE.....	235
PENCARIAN DI MARKETPLACE MEMANFAATKAN ARTIFICIAL INTELLIGENCE	236
PROFILING PENGGUNA MARKETPLACE	238
FORTINET'S LONGSTANDING HISTORY OF AI-DRIVEN SECURITY	241
<i>By. Edwin Lim dan David Finger - Fortinet Indonesia</i>	
PENDAHULUAN	241
MEMPERLUAS PENAWARAN KEAMANAN YANG DIGERAKKAN OLEH AI FORTINET	242
MENGUNAKAN AI UNTUK MERATAKAN BIDANG BERMAIN CYBER.....	243
MEMPREDIKSI MASA DEPAN	243
MENDAPATKAN POSISI ATAS	244
EVOLUSI DAN MASA DEPAN AI	244
BERANGKAT DARI SINI KE SANA	245
AI AS THE ULTIMATE DISRUPTER IN LOGISTICS: HOW TO MANAGE LAST MILE COSTS?	247
<i>By. Rohit Kumar - Rosebay Inc.</i>	
INTRODUCTION	247
AI IN LOGISTICS.....	248

LAST-MILE/FIRST MILE SERVICES OPTIMIZATION VIA VEHICLE ROUTING OPTIMIZATION	249
MAKING A COMPREHENSIVE OPTIMIZATION ROADMAP FOR LOGISTICS COMPANIES	249
CASE STUDY: SAVINGS FROM SINGLE-VEHICLE ROUTE OPTIMIZATION IN BANDUNG AND JAKARTA CITY	250
PLANNING YOUR AI/DIGITAL TRANSFORMATION JOURNEY	251
COMPUTER VISION AT THE FOREFRONT OF AI: WHY INDONESIA SHOULD ACCELERATE ADOPTION AI	253
<i>By. Adhiguna Mahendra dan Kristiyanto - Nodeflux Teknologi Indonesia</i>	
PENDAHULUAN	253
AI IN BRIEF AND OUTLOOK	253
AI COMPUTER VISION: DEFINITION & REAL-WORLD APPLICATIONS	254
GLOBAL AND REGIONAL OUTLOOK ON AI & COMPUTER VISION	255
INDONESIA & AI	257
PEMBAHASAN	257
ANALISIS: REGULASI SEBAGAI FONDASI PENERAPAN AI DI INDONESIA	257
COMPUTER VISION DI INDONESIA: A MOMENTUM FOR DIGITAL ACCELERATION	259
E-KYC & FACE BIOMETRICS:	262
PENTINGNYA PENERAPAN DAN PEMANFAATAN TEKNOLOGI AI COMPUTER VISION DI INDONESIA	264
KESIMPULAN	266
TENTANG PENULIS	267

Kata Pengantar



Kecerdasan buatan atau *Artificial Intelligence* (AI), merupakan salah satu teknologi yang diharapkan dapat memberikan manfaat besar bagi masyarakat, melalui inovasi-inovasi di berbagai sektor industri. Memasuki Era industri 4.0, AI atau teknologi supercanggih ini telah memegang peranan penting dalam perkembangan inovasi dimasa depan. Badan Aksesibilitas Teknologi dan Informasi (BAKTI) Kementerian Komunikasi dan Informatika berkomitmen untuk selalu memfasilitasi stakeholder dalam menumbuhkan kreativitas serta inovasi digital di bidang AI dan Big Data tersebut.

Mengutip pernyataan Presiden Jokowi yang menyebut bahwa Data adalah jenis kekayaan baru, analoginya *the new oil*, bahkan lebih berharga dari minyak. Data yang valid merupakan kunci utama kesuksesan pembangunan sebuah negara, dimana Analisa Data dan AI sangat dibutuhkan untuk memberi masukan, *insight* dan wawasan agar BAKTI Kominfo dapat mengatasi kesejangan digital dan broadband terutama mengentaskan daerah 3T (Terluar, Tertinggal dan Terpencil).

Adapun beberapa sektor yang dapat memanfaatkan teknologi AI diantaranya sektor kesehatan, keuangan, serta pertanian dan lain sebagainya. Khusus untuk peningkatan layanan kesehatan melalui *telemedicine*, juga dilakukan BAKTI dalam mendukung agenda pemerintah dalam Percepatan Transformasi Digital Nasional, sesuai dengan arahan Presiden Jokowi. Pada akhirnya, digitalisasi dan otomatisasi sistem rumah sakit dibutuhkan untuk jalan keluar mengatasi problem kesehatan di tanah air.

Telemedicine merupakan suatu metode pendistribusian layanan dan informasi terkait kesehatan dengan memanfaatkan teknologi komunikasi khususnya AI. *Telemedicine* yang didorong oleh Menristek memungkinkan pasien untuk mendapatkan beragam informasi layanan kesehatan melalui percakapan dengan *chatbot*. Selain itu, pasien juga dapat melakukan registrasi melalui sistem yang dihasilkan oleh teknologi AI.

Khusus dalam situasi pandemi Covid-19, penyelesaian infrastruktur untuk fasilitas layanan kesehatan menjadi prioritas yang utama. Pasalnya, dari 13.011 untuk layanan kesehatan tersebut masih menyisakan 3.126 (titik layanan). Karena tentunya tanpa jaringan internet tersebut maka integrasi data khususnya data Covid ini belum bisa secara agregasi nasional secara lengkap. Ada tersisa dari layanan kesehatan tersebut.

Diharapkan sisa titik layanan kesehatan dapat terjangkau dengan internet secara keseluruhan pada kuartal I tahun 2021, sehingga data pasien Indonesia bisa terintegrasi ke pusat. Apalagi, Kementerian Kesehatan memiliki sebuah program yang namanya *Telemedicine* dan *Telehealth*, namun program tersebut tentunya tidak bisa menjangkau daerah yang belum ada internet.

Pastinya, soal transformasi digital salah satunya menyelesaikan *deployment* akses infrastruktur TIK sampai *the last mile* dimana pelayanan administrasi pemerintahan dan pemukiman masyarakat berada. Jadi untuk mempersiapkan Indonesia, menyelesaikan tugas yang dimandatkan Presiden untuk menghadirkan sinyal 4G di seluruh wilayah tanah air.

Adapun beberapa inisiatif dalam Roadmap Indonesia Digital yang perlu diimplementasikan dalam sektor kesehatan hingga tahun 2024 nanti meliputi (1) Perluasan jangkauan infrastruktur digital dalam mendukung layanan kesehatan melalui *telehealth/telemedicine*, (2) Penerapan registrasi kesehatan digital nasional dalam hal ini manajemen data dan *health record*. Kemudian, (3) Pengembangan hub dan ekosistem teknologi medis, (4) Penerapan *analytics* untuk manajemen penyakit untuk meningkatkan akurasi diagnosa), (5) Perluasan pelacakan kontak *tracing*; dan (6) Implementasi digitalisasi untuk mendorong hidup yang lebih sehat.

Lebih dari itu, pemerintah juga terus melakukan peningkatan literasi digital dan penyiapan SDM atau talenta digital khususnya di bidang AI untuk mendukung pemanfaatan teknologi di bidang kesehatan, dalam hal ini teknologi-teknologi digital telekomunikasi.

BAKTI BANGUN 421 BTS DAN 421 AKSES INTERNET

Tahun depan 2021, BAKTI Kominfo akan membangun 421 BTS baru, sehingga akan terdapat total 542 BTS yang tersebar di 16 kabupaten dan kota. Selain itu, juga akan dibangun akses internet cepat. Dari total 7.652 titik Internet cepat yang dibangun oleh BLU BAKTI, 852 titik berada di Provinsi Nusa Tenggara Timur. 538 titik diantaranya dimanfaatkan untuk sektor pendidikan, sedangkan sisanya dimanfaatkan untuk kantor pemerintahan, pelayanan kesehatan, pusat kegiatan masyarakat, dan sebagainya.

Seyogyanya, perluasan infrastruktur membutuhkan kolaborasi dan sinergitas dan dukungan Pemerintah Daerah, operator telekomunikasi dan pemangku kepentingan. Nah, agar pembangunan BTS dan layanan akses internet gratis berjalan lancar serta tepat guna dan tepat sasaran. Kominfo membutuhkan diantaranya perizinan untuk lahan atau penggelaran kabel untuk BTS, koordinasi dalam hal titik lokasi akses internet, juga kerja sama dalam hal pemanfaatan dan perawatan layanan yang telah dibangun tersebut.

BAKTI DUKUNG TATA KELOLA DATA DENGAN PEMERATAAN INFRASTRUKTUR

Peran BAKTI melalui program *Universal Service Obligation/Kewajiban Pelayanan Universal (KPU/USO)* merupakan bentuk dukungan tidak langsung terhadap perkembangan tata kelola Data dan AI di Indonesia melalui pemerataan infrastruktur telekomunikasi untuk memenuhi target *Sustainable Development Goal (SDG)* di Indonesia.

BAKTI yang berada di bawah Kemkominfo memiliki tugas untuk mengelola dana *universal service obligation (USO)* yang dihimpun dari penyelenggara telekomunikasi sebesar 1,25% dari pendapatan kotor. Adapun dana tersebut digunakan untuk membangun infrastruktur telekomunikasi di Indonesia, termasuk Palapa Ring dan pembangunan menara pemancar (*Base Transceiver Station/BTS*). Namun, dana USO masih belum cukup untuk membangun infrastruktur.

Apalagi, pemerataan infrastruktur telekomunikasi dan informatika yang berkualitas dilakukan dengan cara membangun jaringan *backbone*, *middle-mile*, dan *last-mile*. Mengenai jaringan *middle-mile*, pemerintah terus meningkatkan pembangunan infrastruktur melalui pembangunan jaringan *fiber-link*, *microwave-link*, dan satelit.

Tahun 2023, Indonesia akan meluncurkan satelit multifungsi atau *high-throughput satellite* (HTS) Satria untuk melengkapi lima satelit nasional dan empat satelit asing yang digunakan sekarang. Di samping itu, bersama dengan operator telekomunikasi, BAKTI juga terus mendorong pemerataan jaringan *fixed broadband*. Hal ini juga termasuk optimalisasi penggunaan satelit dengan memanfaatkan teknologi Big Data dan AI untuk penuntasan infrastruktur *last-mile* di wilayah NKRI untuk mendukung program dan komitmen SDG 2030.

Sebagai informasi, SDG lahir pada sidang PBB tentang SDG di Rio de Janeiro (2012) dengan objektives untuk membuat satu *set goal* yang universal untuk memenuhi tantangan yang urgent terkait lingkungan hidup, politik, ekonomi, kesehatan, kesejahteraan yang dihadapi dunia melanjutkan goal dan komitmen dari MDG.

Ketika itu, Indonesia sebagai anggota PBB ikut mendeklarasikan *Millennium Development Goals* (MDGs) sejak pergantian milenium tahun 2000 dan sudah digantikan dengan *Sustainable Development Goals* (SDGs) yang lahir di Rio de Janeiro 2012 mendekati akhir dari 15 tahun perjalanan MDG yang berakhir pada tahun 2015 dengan objektif melahirkan goal yang universal untuk mengatasi urgensi masalah dan tantangan dunia terkait lingkungan, politik, sosial, kemiskinan, ketinggalan (*divide*) dan ekonomi.

Kementerian Kominfo beberapa kali memimpin delegasi ke WSIS (*World Summit on Information Society*) dipimpin oleh Dr. Sofyan Djalil (Geneva, 2003 dan Tunisia, 2005) untuk bersama dunia mencari solusi mengatasi masalah kesenjangan (*divide*) digital, urbanisasi, penetrasi broadband hingga kemiskinan di Indonesia dan di dunia.

Pada waktu MDG Kementerian Kominfo bersama BP3TI, yang sekarang adalah BAKTI Kominfo membuat banyak kajian data daerah 3T dengan berbagai program antara lain PLIK dan MPLIK yang rencananya dapat digelar di ribuan kecamatan, sehingga masalah kesenjangan di rural dan daerah 3T yang rencananya memberikan akses Internet kepada seluruh daerah dengan kategori *blank spot* dan tidak memiliki koneksi. Sayangnya proyek atau upaya besar yang sedang dilakukan itu berjalan ditengah agenda politik Pemilu dan terkena moratorium DPR ketika itu, sehingga implementasinya tidak mencapai target MDG untuk penetrasi Internet di pedesaan.

Di Indonesia masalah kesenjangan digital, rural-urban ditujukan pada daerah 3T termasuk daerah perbatasan; daerah yang tidak memiliki akses jalan darat atau laut dan daerah yang secara ekonomi tidak layak (*feasible*) bagi operator *broadband* komersial. Semoga dengan memanfaatkan Analisa data dan AI maka Indonesia dapat sukses menyelesaikan masalah kesenjangan digital dan desa kecamatan dengan status 3T.

Semoga buku ini bermanfaat dan memberi *insight* bagi banyak perusahaan atau instansi pemerintah yang mulai ancap-ancang untuk mengadopsi AI yang mulai meningkat, khususnya dikalangan data *scientist* di Indonesia. Selain itu, buku ini dapat dijadikan referensi bagi berbagai kalangan akademis dan industri serta menjadi tambahan ilmu untuk kehidupan sehari-hari apalagi kita ketahui bahwa AI juga bisa digunakan untuk menyelesaikan tantangan perusahaan dengan skala besar.

Dengan kata lain, penggunaan teknologi AI ini bisa dimanfaatkan untuk berbagai bidang kehidupan dan juga pemanfaatan teknologi ini menjadi semakin menjamur dimana-mana. Kita pun harus mencontoh negara-negara lain di dunia yang sudah memiliki kemajuan tercepat dalam pengembangan penerapan teknologi ini. Tak hanya itu, dengan membaca buku ini secara mendalam menjadi daya tarik bagi para pengembang teknologi untuk dapat mengembangkan teknologi AI secara maksimal.

Jakarta, Juli 2020

Dirut Badan Aksesibilitas Telekomunikasi dan Informasi (BAKTI)

Ir. Anang Achmad Latif, M.Sc.

Pra Kata



Setelah sukses menerbitkan buku Pertama dengan judul “Big Data Analytics Ecosystems & Solution dengan Apache Hadoop” yang sempat “Best Seller” di Toko Buku Gramedia dan favorit di kalangan masyarakat akhir 2019 serta diluncurkan pada acara DataGovAi 2019 Summit, Expo & Awards bersama Menkominfo (2014-2019) Chief Rudiantara dan para penulis buku. Maka ABDI (Asosiasi Big Data & AI) menerbitkan kembali buku dengan topik “AI & Data Science” dan topik “Data & Cyber Security” dan diluncurkan di acara e-Summit DataGoAi 2020 pada tanggal 24 November Day 1 BRI Hall; 26 November Day 2 BCA Hall; dan 1 Desember 2020 Day 3 BAKTI Hall, untuk mengisi kebutuhan pustaka ilmu pengetahuan dari dan untuk anggota ABDI di sektor AI dan supply chain AI, beberapa diantara para penulis buku tersebut adalah anggota ABDI.

Tahun lalu, pidato kenegaraan Presiden Jokowi di Parlemen (2019) mengingatkan pentingnya kedaulatan data, karena data adalah harta kekayaan baru dan strategis, artinya Bangsa Indonesia harus dapat menguasai teknologi *Big Data* dan memiliki Tata Kelola *Big Data*. Namun dengan semakin maraknya penggunaan AI, ML dan algoritma yang memproses *Big Data* di dunia sebagai sumber daya dan energi untuk menghasilkan *insight* dalam pengambilan keputusan, dimana algoritma AI mempunyai peranan strategis sebagai mesin yang sangat ampuh dan semakin dibutuhkan untuk mengolah data agar lebih efektif dan bermanfaat bagi pembangunan bangsa Indonesia.

Penggunaan AI yang semakin strategis seperti diungkapkan oleh Presiden Rusia bahwa negara yang menguasai dan dapat memanfaatkan teknologi AI berpotensi menguasai dunia, sehingga negara di dunia belomba lomba meluncurkan stranas-nya. Salah satu alasan Menteri Ristek dan Kepala BPPT segera meluncurkan Strategi Nasional (Stranas) AI pada peringatan HUT RI ke 75 dan Presiden Jokowi kembali mengingatkan pada pidatonya tahun ini, 2020 agar para Pejabat Pemerintah harus mengambil keputusan dan kebijakan berdasarkan data dan algoritma yang tepat, bukan berdasarkan intuisi, perasaan dan asumsi. Ilmu *Data Science* hingga AI menjadi penting seperti yang juga pernah ditekankan oleh Mendikbud Nadiem Makarim dan menjadi topik dari buku yang akan diterbitkan oleh ABDI. Inilah alasan mengapa ABDI segera menerbitkan buku tentang “AI & Data Science”.

Editor buku *AI & Data Science* menyusun dan melibatkan berbagai aktor intelektual, pembuat kebijakan dan pakar penulis yang peduli terhadap teknologi AI dan *Data Science* dari manca negara dan berbagai disiplin ilmu dan industri, yang sebageaian besar juga anggota ABDI, sehingga pembaca dapat memperoleh ilmu pengetahuan mengenai AI dan *Data Science* dari berbagai spektrum dan sudut pandang AI dari dalam dan luar negeri sebagai berikut:

Editor: Ketua Umum ABDI – Dr. Rudi Rusdiah, BE. MA.

Kata Sambutan oleh:

1. Menteri Riset dan Teknologi/BRIN – Bambang Brodjonegoro, Ph.D.
2. Kepala Staf Presiden RI – Jend TNI (Purn) Dr. H. Moeldoko

3. Dirut BAKTI Kominfo – Ir. Anang Latief, M.Sc.
4. Kepala BPPT – Dr. Hammam Riza, M.Sc.

Penulis Kontributor :

1. Micronics Internusa PT (bersama ABDI dan Komite.id) – Dr. Rudi Rusdiah, BE. MA (CEO)
2. Indian Institute of Information Technology (IIIT) Allahabad, India - Dr. Sonali Agarwal, Ph.D.; University Budi Luhur (UBL) Dr. Ir. Wendi Usino, M.Sc.
3. BRI – Tio Anta Wibawa, B.Sc. (Data Science) dan Dr. Indra Utoyo (Director)
4. BCA – Hendra Lembong (Director) dan Ir. Jayaprawirya Diah (EVP)
5. BPPT – Dr. Hammam Riza (Kepala BPPT) dan Dr. Hary Budiarto (Staf Ahli)
6. GDP Venture, GDP Labs, Kaskus – On Lee (CEO, CTO)
7. Prosa.AI, Head AI Centre ITB – Dr. Ayu Purwarianti, ST, MT (Chief AI Scientist & Co-Founder)
8. Coqniyx – Frado Sibarani, M.Sc, MBA (Chief AI Officer)
9. University Budi Luhur - Irwan Susanto, Suwardiman, Ummu Habibah
10. University Budi Luhur - Amin Muzaeni, Khoirunnisa, Yogi Agung Subekhti, M Windy Trihantoro
11. Rosebay Inc, India – Rohit Kumar Founder (CEO)
12. Cakra.AI - Rizaldi Sistiabudi, Ph.D (CEO) dan Angga Airlangga, M.Kom (CTO)
13. Ucoach - Dr. Franz Budi Pratama, SE, MBA dan Gendro Salim (Founder)
14. Potensiana.com - Febi M Faizal (Founder & CTO); Qori Utama Co Founder dan Dr. Franz Budi Pratama, SE, MBA
15. NodeFlux – Adhiguna Mahendra, M.Sc, Ph.D (Chief AI Reserach Officer & Product Innovation)
16. Wardaya College - Dr. Anton Wardaya, M.Sc. (Founder)
17. Aimy Robotics & Global Management Group – Dr. Peng Chan dan Dr. Will Smith

Buku AI dan Data Science ini akan bermanfaat bagi:

- *Data Science, Data Engineer, AI Master* yang mulai menerapkan AI dan ML
- Siswa yang ingin belajar teknologi AI dan Data Science mengingat AI dan Data Science sebagai profesi yang sangat dicari dan high demand globally.
- Profesional dan komunitas IT (*Information Technology*) dan *Cyberspace* dari berbagai bidang karena AI dimanfaatkan oleh hampir semua interdisipline dan bidang. Setelah Big Data, kini AI sebagai komoditas masa depan bagi semua profesi, masyarakat maupun semua bidang industri dan perdagangan.
- Stakeholder pembuat dan implementor rancangan Regulasi terkait etika menerapkan antara lain AI, *Big Data* dan ML

PENGHARGAAN (ACKNOWLEDGMENT)

Buku ini tidak mungkin dapat diterbitkan tanpa dukungan dari banyak pihak dan staholder. Terima kasih kepada para Penulis Kata Sambutan yang terhormat, Jenderal TNI (Purn) Dr. Moeldoko, Prof. Bambang Brodjonegoro, Ph.D, Ir. Anang Latief dan Dr. Hammam Riza yang berkenan membuat kata sambutan dan memberi semangat dan dukungan kepada para penulis manca negara.

Para Pakar Akademisi antara lain Indian Institute of Information Technology (IIIT) Allahabad, Universitas Budi Luhur (UBL) dan Wardaya College.

Para Pakar dari industri dan komunitas antara lain: Micronics Internusa (Accessindo Internusa), BRI, BCA, GDP Venture, Prosa.AI, Coqnixy, Rosebay Inc, Cakra.AI, Ucoach, Potensiana, Nodeflux, Aimy Robotics.

Juga apresiasi kepada tim pendukung, editing, layout dan proses penerbitan buku ini oleh Samuel dan Deden; kegiatan administrasi dan logistik oleh Budi, Ida Mindasari, Yuliana, Angela, Giok, Ros, Yuniarti, Theresia; masalah teknis hardware & pelatihan oleh FX Winarto, Jorry, Kasmirus, Arifin, Sastrio dan tim engineer Micronics Group. Tim majalah Komite.id Juanda dan Nurul yang ikut membantu promosi buku ini.

Semoga buku ini dapat memberikan kontribusi yang besar bagi dunia ilmu pengetahuan dan Pustaka Nasional dalam bidang AI dan *Data Science* di Indonesia.

Jakarta, Juli 2020
Editor

Dr. Rudi Rusdiah, BE, MA.



Targeting Class Imbalance Problem in Credit Card Fraud Detection using Generative Adversarial Network Synthetic Oversampling Framework

Sonali Agarwal, Narinder S. Punn, Sanjay K. Sonbhadra and Wendi Usino
Indian Institute of Information Technology Allahabad and UBL

Abstract

With the growing phase of digitization and credit card transactions in the financial industry, the fraudulent transaction cases have gained a similar spike causing a huge amount of financial loss. From the daily average pool of billions of transactions, it is quite challenging to identify the transactions as a fraud case due to its relatively less frequent occurrence than non-fraud cases. The Kaggle Credit Card Fraud Detection (CCFD) challenge 2018 offers the credit card transactions information that consists of 0.172% of fraud cases resulting. Due to this, machine learning algorithms struggle to discover the patterns associated with the minority class (fraud cases). Inspired from this problem, an efficient Generative Adversarial Network Synthetic Oversampling Framework (GANSOF) is proposed to oversample the dataset with an equal ratio of fraud and non-fraud cases forming the balanced dataset. The quality of generated data is evaluated using the Jensen-Shannon (JS) divergence metric that quantifies the similarity between the two probability distributions. The balanced dataset is then used to train the following machine learning classifiers: Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest (RF). The models were evaluated on the hold-out validation set from real samples where SVM and RF produced the best results to detect the fraud and non-fraud transactions. The results of the proposed approach show significant improvement in the performance of the classification models measured in terms of precision, recall, accuracy and F1-score.

Keywords: *Classification, Data imbalance, Fraud transaction, Generative adversarial network, Synthetic data sampling.*

INTRODUCTION

The continuous expansion of digitization is bringing revolution in the financial industry. In recent years, there has been tremendous increase in the credit card transactions along with the rampant of the fraud transactions in both online and offline payment platforms. The fraud transactions resemble the unfair means of the payments performed from the illegal access of the credit card details causing huge loss to the customers and financial economy. These criminal activities necessitate the efficient detection of fraudulent cases in the financial sector.

Every day billions of transactions are made from online and offline payment gateways, which may consist of marginal fraudulent transactions due to which identifying such transactions from the huge pool of imbalanced data is certainly a challenging and difficult task. The high level of skewness in data hinders the training process of the classification models from learning the characteristic features of the fraud transactions.

Imbalanced Learning

The imbalanced learning problem^{1,2} occurs in the presence of unbalanced distribution of data samples where the classes having significantly higher number of sam-ples are referred as majority classes, as compared to other ill-defined classes called minority classes. This problem further extends across various real-world applications^{3, 4, 5} such as medical diagnosis, fraud transaction detection, and other target-specific learning problems⁶, etc. The Kaggle Credit Card Fraud Detection (CCFD) 2018 challenge⁷ offers such highly imbalanced data that consists of 0.172% of fraud transaction from the approximate pool of 3 lakhs transactions. In order to enable learning from such skewed data, requires some data processing technique that restores the balance among classes in the dataset. This can either be achieved by the under-sampling or the over-sampling approaches.⁸ In under-sampling techniques,

1 Longadge, Rushi, and Snehalata Dongre. "Class imbalance problem in data mining review." arXiv preprint arXiv:1305.1707 (2013).

2 Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanue, and Gong Bing. "Learning from class-imbalanced data: Review of methods and applications." *Expert Sys-tems with Applications* 73 (2017): 220-239.

3 Punn, Narinder Singh, and Sonali Agarwal. "Inception U-Net Architecture for Semantic Segmentation to Identify Nuclei in Microscopy Cell Images." *ACM Transactions on Multi-media Computing, Communications, and Applications (TOMM)* 16, no. 1 (2020): 1-15.

4 Gangwar, Akhilesh Kumar, and Vadlamani Ravi. "WiP: Generative Adversarial Network for Oversampling Data in Credit Card Fraud Detection." In *International Conference on In-formation Systems Security*, pp. 123-134. Springer, Cham, 2019.

5 Taha, Altyeb Altaher, and Sharaf Jameel Malebary. "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine." *IEEE Access* 8 (2020): 25579-25587.

6 Punn, Narinder Singh, and Sonali Agarwal. "Testing Concept Drift Detection Technique on Data Stream." In *International Conference on Big Data Analytics*, pp. 89-99. Springer, Cham, 2018.

7 Kaggle. Credit Card Fraud Detection. 2018. Retrieved February 6, 2020 from <https://www.kaggle.com/mlg-ulb/creditcardfraud>

8 Yap, Bee Wah, Khatijahusna Abd Rani, and et al. "An application of oversampling, under-sampling, bagging and boosting in handling imbalanced datasets." In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)*, pp. 13-22. Springer, Singapore, 2014.

the intention is to develop a sub-sampled dataset from the original dataset via removing certain samples belonging to the majority class. Whereas, the over-sampling techniques tend to generate the artificial samples of the minority class to reduce the skewness in the class distribution. The resulting balanced dataset becomes suitable for classification algorithms to discover the feature patterns of the fraudulent trans-actions for robust and scalable learning.

Overview

In this article, an efficient Generative Adversarial Network Synthetic Oversampling Framework (GANSOF) is proposed that utilizes the state-of-the-art generative adversarial network⁹ approach and synthetic minority oversampling technique¹⁰ to target the skewness in the credit card transactions. The two adversarial networks: generator and discriminator that follow the Multi Layer Perceptron (MLP) architecture, are trained on 70% of the original samples, whereas remaining samples are kept as hold-out set for final evaluation. The generator network outputs the synthetic data and discriminator supervises the quality of synthetic data with that of real or original data. Jensen-Shannon (JS) divergence¹¹ measure is utilized which is based on Kull-back-Leibler (KL) divergence¹², to verify that the samples produced by generator are similar to the real samples. From several trials, it was observed that the GAN-based approach produced quite significant results in over-sampling the minority classes for generating balanced dataset. Furthermore, the classification results evaluated on the hold-out validation set of real samples, are best produced on the balanced dataset generated via proposed GANSOF approach and that too by using Support Vector Machine (SVM) and Random Forest (RF) classifiers out of other techniques: Logistic regression (LR), K-Nearest Neighbours (KNN). Fig. 1 illustrates the overall idea of the proposed framework.

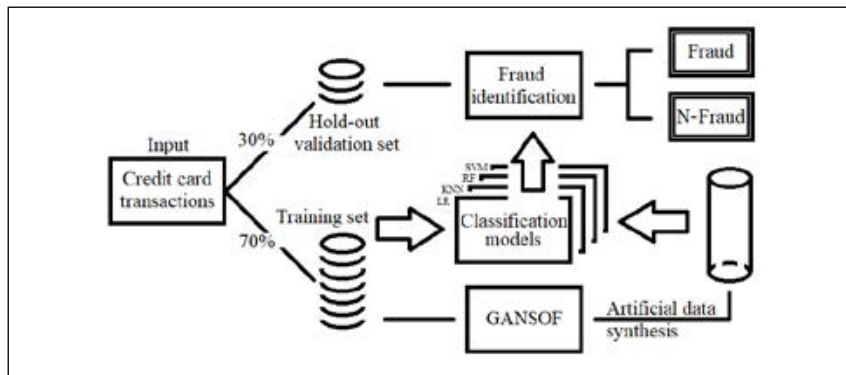


Figure 1

Schematic representation of overall framework for credit card fraud detection.

9 Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Advances in neural information processing systems*, pp. 2672-2680. 2014.

10 Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

11 Goodfellow, Ian, Jean Pouget-Abadie, cs., *Loc. Cit.*

12 *Ibid.*

RELATED WORK

The continuously growing credit card fraudulent transactions have gained significant importance and due to its challenging characteristics, where very few are fraudulent cases, draws more attention from the research community. The objective here is to eliminate the skewness in the imbalanced data to better understand the fraudulent financial statements and to reduce the number of false predictions. With this concern, many approaches^{13, 14, 15} have been proposed that exploit the skewness problem at the algorithm level, data level, and data-algorithm (hybrid) level.

The algorithm driven approaches are cost-sensitive which optimize the model weight matrix to adapt the skewed distribution of data. Mostly, the algorithm driven approaches are based on the variations of Extreme Learning Machines (ELM)¹⁶ that is a Single Layer Feedforward Neural Network (SLFN). The Weighted-ELM (W-ELM)¹⁷ is one among the variants of ELM that targets on learning from imbalanced information. W-ELM produces better results on the imbalanced dataset by assigning different weights among the available classes with the subject of prioritizing the learning from minority classes. However, the shallow structure of W-ELM restricts to learn the complex patterns. Later, Hierarchical-ELM (H-ELM) based on multilayer perceptron network was proposed by Han et al.¹⁸ to improve upon the feature representation and the classification performance. Whereas the data level approaches^{19, 20} focus on under-sampling the dataset by removing the samples of majority class or over-sampling the dataset by synthesizing the minority class in order to eliminate the skewness in the class distribution. These approaches are easier to understand and implement when compared to other cost-sensitive approaches.²¹ The under-sampling approaches follow certain procedures to keep, delete, and combination of keep-delete samples belonging to the majority class. However, the discarded samples may have some valuable information, hence the approach could lead to biased learning. The oversampling techniques generate new realistic artificial samples of the minority class. The SMOTE²² is one of the typical methods that utilize the statistical feature set, based on the nearest neighbour to generate the synthetic samples of minority class rather than just duplicating the information, and

13 Guo HX, Li YJ, Shang J. Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl.* 2017;73:220-39.

14 Lee, Wonji, Chi-Hyuck Jun, and Jong-Seok Lee. "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification." *Information Sciences* 381 (2017): 92-103.

15 Wu, Zhenyu, Wenfang Lin, and Yang Ji. "An integrated ensemble learning model for imbalanced fault diagnostics and prognostics." *IEEE Access* 6 (2018): 8394-8402.

16 Salaken, Syed Moshfeq, Abbas Khosravi, Thanh Nguyen, and Saeid Nahavandi. "Extreme learning machine based transfer learning algorithms: A survey." *Neurocomputing* 267 (2017): 516-524.

17 Zong, Weiwei, Guang-Bin Huang, and Yiqiang Chen. "Weighted extreme learning machine for imbalance learning." *Neurocomputing* 101 (2013): 229-242.

18 Han, Hong-Gui, Li-Dan Wang, and Jun-Fei Qiao. "Hierarchical extreme learning machine for feedforward neural network." *Neurocomputing* 128 (2014): 128-135.

19 Galar M, Fernandez A, Barrenechea E. EUSBoost: enhancing ensembles for highly imbalanced datasets by evolutionary undersampling. *Pattern Recogn.* 2013;46(12):3460-71.

20 Castellanos, Francisco J., Jose J. Valero-Mas, Jorge Calvo-Zaragoza, and Juan R. Rico-Juan. "Oversampling imbalanced data in the string space." *Pattern Recognition Letters* 103 (2018): 32-38.

21 Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks* 106 (2018): 249-259.

22 Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. *Loc. Cit.*

thereby alleviates the overfitting problem without the loss of information. Despite this, SMOTE is viable to cause class overlapping issue due to its negligence of class of neighbouring examples. It also fails to accommodate the characteristic features of the original distribution of high dimensional data.²³

In this article, an efficient GANSOF approach based on the GAN and SMOTE is proposed that tackles the class imbalance problem by generating the artificial data samples of minority class, while also preserving the original sample class distribution. The generated data is then utilized using machine learning algorithms to effectively detect fraudulent transactions.

PROPOSED ARCHITECTURE AND METHODOLOGY

The GANSOF is comprised of GAN and SMOTE approaches to oversample the data. The GAN's generator (G) and discriminator (D) adversarial networks are trained to synthesize the artificial minority data samples. Table 1 describes the layered architectures of generator and discriminator neural networks with total trainable and non-trainable parameters of 1.3 M and 5.6 K.

Network	Layer	Activation	Dropout	Output	Parameters
Generator	Input	–	–	30x100	–
	Dense – 1	Leaky ReLU	–	256	768256
	Batch Norm. – 1	–	–	256	1024*
	Dense – 2	Leaky ReLU	0.3	512	131584
	Batch Norm. – 2	–	–	512	2048*
	Dense – 3	Leaky ReLU	0.3	512	262656
	Batch Norm. – 3	–	–	512	2048*
	Dense – 4	Leaky ReLU	0.2	128	65664
	Batch Norm. – 4	–	–	128	512*
	Dense – 5	tanh	–	30	3870
Discrim.	Dense – 6	Leaky ReLU	–	512	15872
	Dense – 7	Leaky ReLU	–	128	65664
	Dense – 8	Leaky ReLU	–	32	4128
	Dense – 9	Sigmoid	–	1	33

*indicates non-trainable parameters, input layer has 30 features with a batch of 100 samples

Table 1

Layer architecture of generator and discriminator.

The networks (G and D) consist of dense layers that are activated using leaky ReLU^{24, 25} function except for the output layers which are activated using tanh and sigmoid functions in the generator and discriminator networks respectively. The dense layers are accompanied by batch normalization to

23 Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

24 Karlik, Bekir, and A. Vehbi Olgac. "Performance analysis of various activation functions in generalized MLP architectures of neural networks." International Journal of Artificial Intelligence and Expert Systems 1, no. 4 (2011): 111-122.

25 Punn, Narinder Singh, and Sonali Agarwal, *Loc. Cit.*

avoid the covariance shift problem²⁶ and achieve better results.

With the extensive experiments, it was observed that the generated artificial data using GAN alone can further be improved to produce robust data. Therefore, the trained discriminator is then utilized with the SMOTE approach to generate the data samples. The samples generated from SMOTE are supervised by the discriminator to filter-out the generated outliers. The final oversampled dataset is generated by mixing the samples from both the GAN and discriminator supervised SMOTE approaches.

The synthetic balanced dataset is utilized for training and validation of the classification models. The classification results are evaluated on the hold-out set from the original sample distribution using precision, recall, accuracy and F1-score defined based on the confusion matrix.²⁷

$$\text{Precision, } P = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall, } R = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Accuracy, } A = \frac{TP + TN}{TP + FN + FP + FN} \quad (3)$$

$$F1 = \frac{2PR}{P + R} \quad (4)$$

where True Positive (TP) and True Negative (TN) indicates the number of transactions correctly classified as fraud and non-fraud respectively, and False Positive (FP) and False Negative (FN) indicates the number of transactions incorrectly classified as fraud, and non-fraud respectively.

Synthetic Data Generation Using GANSOF

The vanilla GAN [9] is a deep neural network consisting of two following adversarial networks: generator (G) and discriminator (D). During the training phase, the G network is fed with randomly initialized noise vector z to synthesize sample as given in equation 5. Whereas for some sample x , D network focuses on predicting the probability that whether it is real or synthetic (equation 6).

$$G_{\theta}(z) = Z \quad (5)$$

$$D_{\phi}(x) = \begin{cases} 1, & \text{if } x \in X \\ 0, & \text{if } x \in Z \end{cases} \quad (6)$$

where X indicates the set of real samples, Z is the set of artificially generated samples, θ and ϕ indicates the distinct sets of $(\omega_0^0, \dots, \omega_{n_0}^0, \omega_0^1, \dots, \omega_{n_1}^1, \dots, \omega_{n_m}^m)$ trainable parameters as $\omega_{n_m}^m$ indicating weight associated with m^{th} layer and n^{th} neuron.

The D and G networks are trained to maximize the probability function

²⁶ Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

²⁷ Mekterović, Igor, Ljiljana Brkić, and M. I. R. T. A. Baranović. "A systematic review of data mining approaches to credit card fraud detection." WSEAS Transactions on Business and Economics 15 (2018): 437.

$\log(D_\phi(x))$ and to minimize the $\log(1 - D_\phi(G_\theta(z)))$ respectively. This min-max value function $V(G,D)$ over some sample x is represented in equation 7, where \mathbb{E}_x and \mathbb{E}_z indicates the expected value over the real instances and generator inputs respectively.

$$\min_G, \max_D V(D, G) = \mathbb{E}_x \log D_\phi(x) + \mathbb{E}_z \log(1 - D_\phi(G_\theta(z))) \quad (7)$$

Fig. 2 illustrates the training process along with the flow of information in GAN ar-chitecture for a single sample. The training halts when discriminator is unable to dis-tinguish between real and synthetic samples.

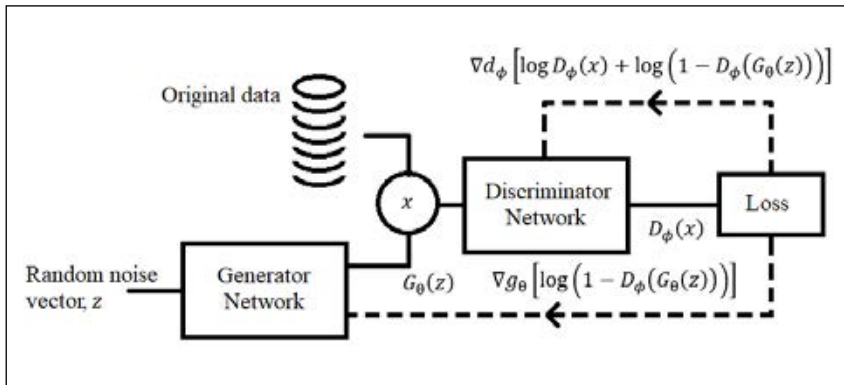


Figure 2
Schematic representation of the training process for a single sample in GANs.

SMOTE is an oversampling technique that creates new synthetic observations. Initially, it identifies the feature vector and nearest neighbours in the minority class. Then the difference is computed between the considered instance and the feature vector of the nearest neighbour, which is multiplied with some random number between 0 and 1. The novel sample is identified by adding this value to the feature vector of the considered sample. The final balanced dataset is generated by aggregating the samples from GAN and SMOTE under the supervision of trained discriminator that is evaluated by JS metric (based on KL) as represented in equation 8 and 9. It helps to better quantify the difference or similarity between the two probabilistic distributions. The JS value lies between 0 and 1, with 0 being the identical

$$JS(Z||X) == JS(X||Z) = \frac{1}{2} KL\left(X||\frac{X+Z}{2}\right) + \frac{1}{2} KL\left(Z||\frac{X+Z}{2}\right) \quad (8)$$

$$KL(P||Q) = \sum_{f \in \mathcal{F}} P(f) \log\left(\frac{P(f)}{Q(f)}\right) \quad (9)$$

case and 1 being maximum divergence.

where $||$ indicates divergence operation, P and Q are any arbitrary probability distribution defined on common probability space \mathcal{F} .

Credit Card Fraud Detection

The CCFD 2018 challenge²⁸ of credit card transactions, aims at binary classification problem to identify and distinguish the two classes of transactions: fraud (1) and non-fraud (0). After generating the resampled data from the original data, the machine learning classifiers are trained to analyze and discover the pattern associated with fraud detection. In this article, following state-of-the-art machine learning algorithms are proposed to utilize for classification: LR, KNN, SVM, and DT using machine learning python library sklearn. These classifiers are trained using the weighted average of binary cross entropy and hinge loss²⁹ as the convex optimization functions. Finally, the trained models are evaluated on the hold-out validation set from the real data samples.

EXPERIMENTATION AND RESULTS

Dataset

Kaggle CCFD 2018 challenge³⁰ offers the numerically valued dataset consisting of 284,807 transactions recorded for 2 days in September 2013, executed by European cardholders. The dataset is highly skewed comprising of 492 transactions marked as the fraudulent cases. The non-fraud transactions are marked with class label 0 whereas fraud transactions are marked with 1.

Sharing such financial information publicly involves high security risks and privacy concerns due to which original features are transformed using Principle Component Analysis (PCA).³¹ Principal components (V_1, V_2, \dots, V_{28}) along with time and amount columns are represented as a feature set in CCFD challenge. The dataset features were already scaled except the time and amount that are scaled using robust scalar from sklearn, in order to have amount and time sensitive training. The final dataset was prepared with the 30 input features and 1 target feature.

Training and Evaluation

GANSOF. The CCFD dataset is split into 70% as the training set and remaining 30% for validation, where training set and test consists of 0.169% and 0.181% of fraud transactions respectively. The training set along with the random noise is utilized to train the adversarial networks. To facilitate the training process, Adam optimizer³² is utilized which is among the best optimizer techniques to update the training parameters. The convergence of GAN and JS with respect to each iteration is illustrated in fig. 3. The JS metric value was further improved by adding fraud samples from SMOTE

²⁸ Kaggle, *Loc. Cit.*

²⁹ Lapin, Maksim, Matthias Hein, and Bernt Schiele. "Loss functions for top-k error: Analysis and insights." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1468-1477. 2016.

³⁰ Kaggle, *Loc. Cit.*

³¹ Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2, no. 4 (2010): 433-459.

³² Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

under the supervision of trained discriminator making 1:1 ratio of fraud and non-fraud samples.

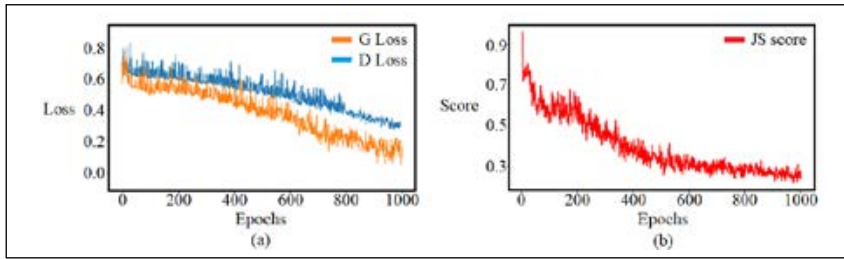


Figure 3

(a) Loss convergence for G and D networks and (b) JS divergence score per epoch.

In the final dataset, 70% of the minority samples are synthesized using GAN and the remaining 30% from discriminator supervised SMOTE as highlighted in table 2. Though there is a significant improvement in JS score, however, ablation study is conducted to illustrate the impact of mixing the GAN and supervised SMOTE approaches in the results and discussion section.

Data source	Approach	Fraud cases (1)	Non-fraud cases(0)
Original	–	492	284315
GANSOF	GAN	198676	–
	Supervised SMOTE	85147	–
Total samples		284315	284315

Table 2

Composition of original and generated data.

Dimensionality reduction and clustering. The quality of synthetic data can also be verified through visualization. The high dimensional feature set is reduced by employing dimensionality reduction techniques³³ such as t-SNE. Fig. 4 shows the clusters generated using t-SNE from original sub-samples consisting of randomly selected set of 1172 samples with 170/1002 fraud/non-fraud samples, and generated sub-samples consisting of randomly selected set of 1100/1100 fraud/non-fraud samples. The visual representation indicates that the generated samples can be made to fit on classification models. The classification models are evaluated with precision, recall, accuracy and F1-score on the real samples. The model with high precision and recall value has low false acceptance and rejection rate, indicating that the classifier model is robust to recognize the positive fraudulent samples.

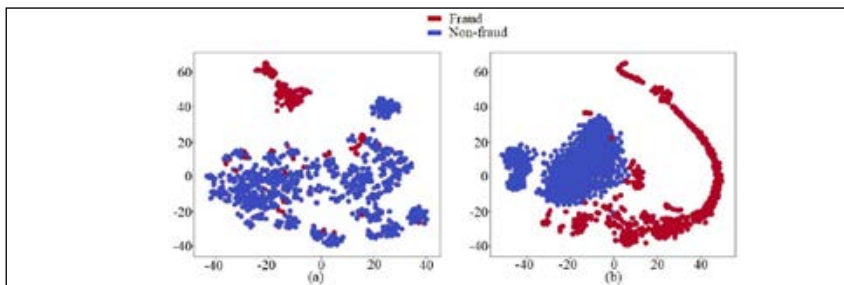



Figure 4

t-SNE 2D representation of (a) Original and (b) GANSOF generated samples.

33 Van Der Maaten, Laurens, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative." J Mach Learn Res 10, no. 66-71 (2009): 13.

Results and Discussion

As highlighted in table 3, the classification models: LR, KNN, SVM, and RF trained on GANSOF generated samples, performed better in classifying the fraud transaction than on classifying it directly on the original, GAN and SMOTE generated data. It was also found that SVM and RF performed better in terms of precision, recall, accuracy and F1-score for identifying the fraud transactions than KNN and LR. It is observed from statistics in table 3 that accuracy metric is not fit to evaluate the performance of the proposed approach due to which precision and recall metrics are compared, to verdict SVM and RF as best models to detect the fraudulent transactions.

 **Table 3** —
Fraud identification results with different data sources and classifiers.

Data source	Fraud cases %	Classifier	Precision	Recall	Accuracy	F1-score
Original	0.172	LR	0.03	0.02	0.97	0.02
		KNN	0.01	0.02	0.96	0.01
		SVM	0.04	0.03	0.97	0.03
		RF	0.03	0.03	0.97	0.03
SMOTE	50	LR	0.92	0.47	0.96	0.62
		KNN	0.91	0.47	0.95	0.62
		SVM	0.92	0.48	0.96	0.63
		RF	0.92	0.48	0.96	0.63
GAN	50	LR	0.93	0.50	0.97	0.65
		KNN	0.93	0.45	0.97	0.61
		SVM	0.93	0.49	0.98	0.64
		RF	0.93	0.50	0.96	0.65
Supervised SMOTE	50	LR	0.92	0.51	0.95	0.66
		KNN	0.92	0.45	0.97	0.60
		SVM	0.92	0.51	0.97	0.66
		RF	0.93	0.52	0.97	0.67
GANSOF	50	LR	0.95	0.59	0.97	0.73
		KNN	0.97	0.58	0.96	0.73
		SVM	0.97	0.61	0.99	0.75
		RF	0.97	0.60	0.97	0.74

CONCLUSION

In this article, the challenging problem of heavy class imbalance is resolved by efficiently oversampling the minority fraudulent cases using the GAN-SMOTE based procedure named as GANSOF. The quality of generated samples is visualized with two-dimensional feature space using t-SNE dimensionality reduction technique. The classification models are trained on the artificially synthesized samples and evaluated on the original samples set from the original data. With extensive experiments, it was observed that the SVM and RF performed better among the other discussed classifiers. Not limited to this work, the same approach can be extended to complex datasets

from other domains and more trials can be conducted with other GAN based hybrid architectures to accommodate more challenging datasets for better results.



REFERENCES

Longadge, Rushi, and Snehalata Dongre. "Class imbalance problem in data mining review." arXiv preprint arXiv:1305.1707 (2013).

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. "Learning from class-imbalanced data: Review of methods and applications." *Expert Systems with Applications* 73 (2017): 220-239.

Punn, Narinder Singh, and Sonali Agarwal. "Inception U-Net Architecture for Semantic Segmentation to Identify Nuclei in Microscopy Cell Images." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, no. 1 (2020): 1-15.

Gangwar, Akhilesh Kumar, and Vadlamani Ravi. "WiP: Generative Adversarial Network for Oversampling Data in Credit Card Fraud Detection." In *International Conference on Information Systems Security*, pp. 123-134. Springer, Cham, 2019.

Taha, Altyeb Altaher, and Sharaf Jameel Malebary. "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine." *IEEE Access* 8 (2020): 25579-25587.

Punn, Narinder Singh, and Sonali Agarwal. "Testing Concept Drift Detection Technique on Data Stream." In *International Conference on Big Data Analytics*, pp. 89-99. Springer, Cham, 2018.

Kaggle. Credit Card Fraud Detection. 2018. Retrieved February 6, 2020 from <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Yap, Bee Wah, Khatijahusna Abd Rani, and et al. "An application of oversampling, under-sampling, bagging and boosting in handling imbalanced datasets." In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)*, pp. 13-22. Springer, Singapore, 2014.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Advances in neural information processing systems*, pp. 2672-2680. 2014.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

Guo HX, Li YJ, Shang J. Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl.* 2017;73:220-39

Lee, Wonji, Chi-Hyuck Jun, and Jong-Seok Lee. "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification." *Information Sciences* 381 (2017): 92-103.

Wu, Zhenyu, Wenfang Lin, and Yang Ji. "An integrated ensemble learning model for imbalanced fault diagnostics and prognostics." *IEEE Access* 6 (2018): 8394-8402.

Salaken, Syed Moshfeq, Abbas Khosravi, Thanh Nguyen, and Saeid Nahavandi. "Extreme learning machine based transfer learning algorithms: A survey." *Neurocomputing* 267 (2017): 516-524.

Zong, Weiwei, Guang-Bin Huang, and Yiqiang Chen. "Weighted extreme learning machine for imbalance learning." *Neurocomputing* 101 (2013): 229-242.

Han, Hong-Gui, Li-Dan Wang, and Jun-Fei Qiao. "Hierarchical extreme learning machine for feedforward neural network." *Neurocomputing* 128 (2014): 128-135.

Galar M, Fernandez A, Barrenechea E. EUSBoost: enhancing ensembles for highly imbalanced datasets by evolutionary undersampling. *Pattern Recogn.* 2013;46(12):3460-71

Castellanos, Francisco J., Jose J. Valero-Mas, Jorge Calvo-Zaragoza, and Juan R. Rico-Juan. "Oversampling imbalanced data in the string space." *Pattern Recognition Letters* 103 (2018): 32-38.

Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks* 106 (2018): 249-259.

Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

Karlik, Bekir, and A. Vehbi Olgac. "Performance analysis of various activation functions in generalized MLP architectures of neural networks." *International Journal of Artificial Intelligence and Expert Systems* 1, no. 4 (2011): 111-122.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Mekterovi, Igor, Ljiljana Brki, and M. I. R. T. A. Baranovi. "A systematic review of data mining approaches to credit card fraud detection." *WSEAS Transactions on Business and Economics* 15 (2018): 437.

Lapin, Maksim, Matthias Hein, and Bernt Schiele. "Loss functions for top-k error: Analysis and insights." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1468-1477. 2016.

Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2, no. 4 (2010): 433-459.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Van Der Maaten, Laurens, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative." *J Mach Learn Res* 10, no. 66-71 (2009): 13.