

Machine Learning-based Models for Disease Prediction

Dr. Eng. Norma Latif Fitriyani, S.Kom., M.IM.

Department of Artificial Intelligence and Data Science
Sejong University, Seoul, Republic of Korea
Faculty page: <https://home.sejong.ac.kr/~norma>



**Norma Latif
Fitriyani**



Research interest: Health Informatics, Machine Learning, Deep Learning, Artificial Intelligence (AI), Image Processing, and Data Analytics

Email: norma@sejong.ac.kr

Contents

1. Introduction
2. Machine Learning
3. Machine Learning-based Disease Prediction Models
4. Machine Learning-based Model for Disease Prediction Applications
5. Conclusion



LEARNING

01

Introduction

MS | LEARNING



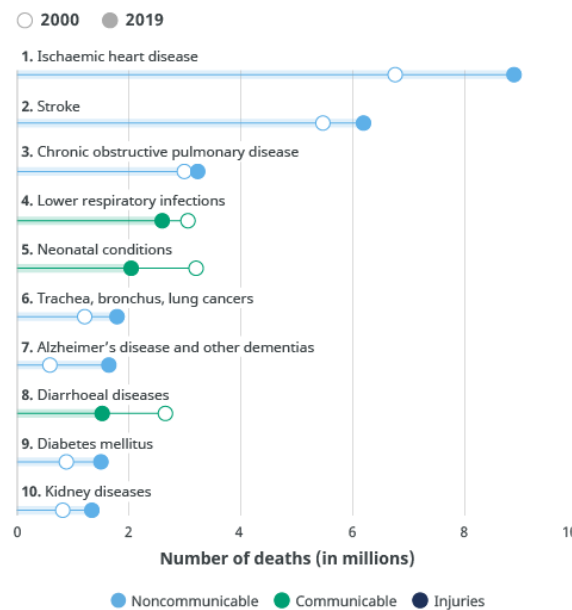
세종대학교
SEJONG UNIVERSITY

ITATS
INSTITUT
TEKNOLOGI
ADHI TAMA
SURABAYA

Introduction

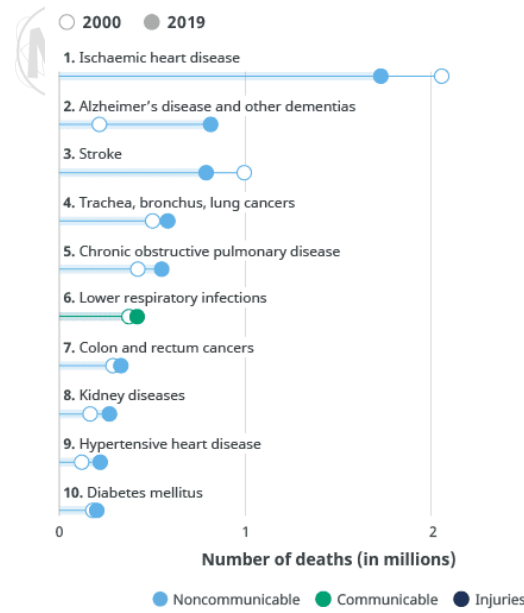
- According to World Health Organization report [1], in 2019, the top 10 causes of death accounted for 55% of the 55.4 million deaths worldwide.
- These top 10 serious diseases take an ‘immense and increasing toll on lives’ [1].

Leading causes of death globally



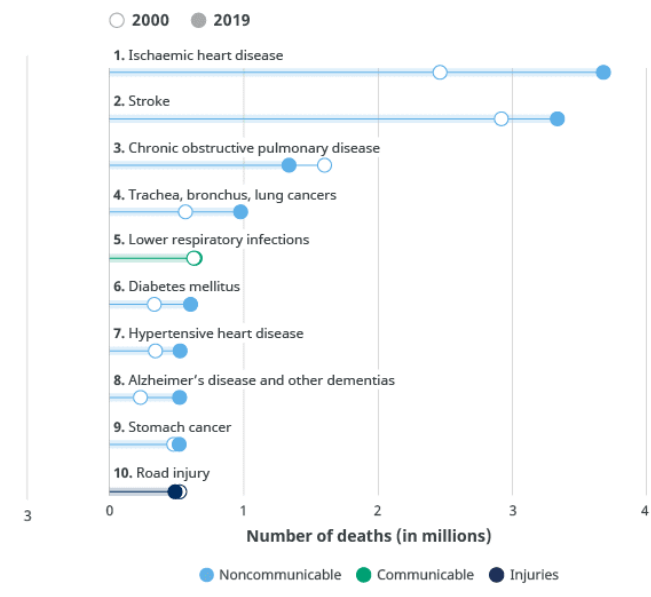
(a)

Leading causes of death in high-income countries



(b)

Leading causes of death in upper-middle-income countries



(c)

Fig 1. Worldwide leading cause death (a), in high-income countries (b), in upper-middle-income (c) [1]

Introduction

- As the number of deaths due to chronic diseases rose annually, the cost of medical diagnosis, tests, and treatment also followed rising [2, 3] .
- Several methods and strategies should be developed as a solution to help individuals more easily and cost effectively check their health status, thus could help early detect the diseases and prevent from the occurrence of the worst-case scenario.

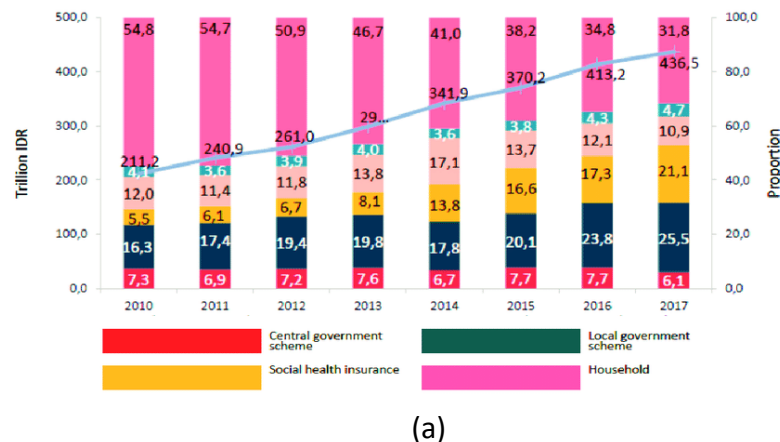


Fig 2. Total Health Expenditure of Indonesia (a), USA (b), Korea (c). [4, 5, 6, 7]

Figure 1. Cumulative Hospital Expense Growth is More Than Double the Cumulative Increases in Medicare IPPS Reimbursement, 2019-2022

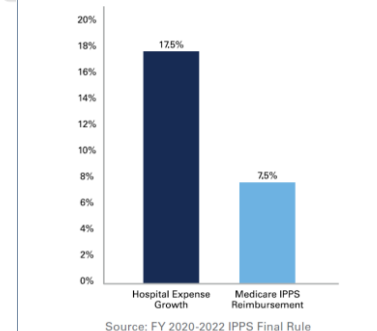
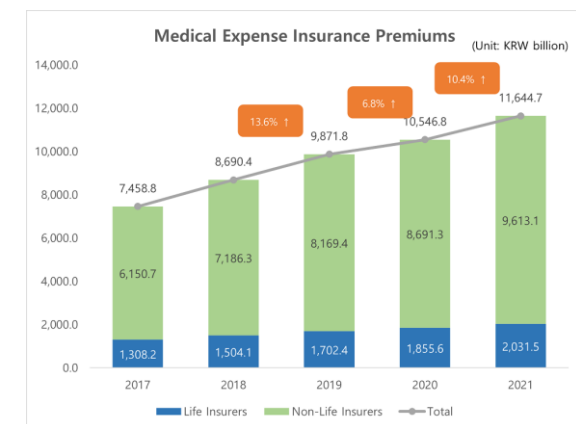
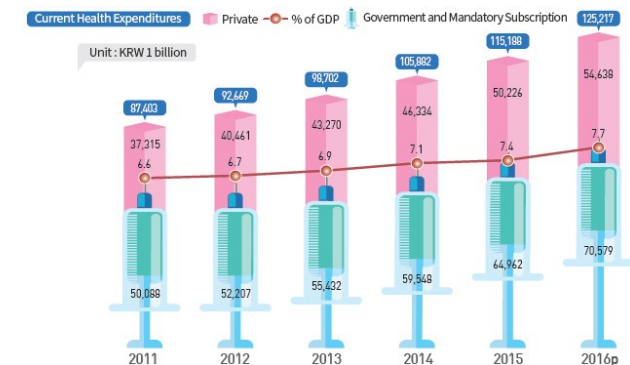
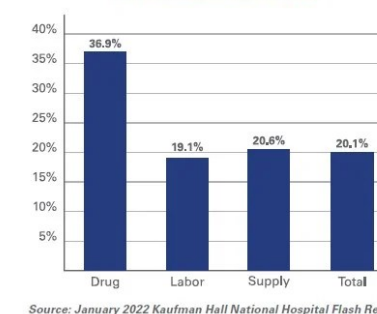


Figure #1: Increase in Hospital Expenses Per Patient from 2019 to 2021



(b)

(c)

Introduction

- One of the solutions that could be used to early detect the disease is machine learning-based prediction models development and utilization.
- Recent studies have utilized machine learning algorithms as decision-making tools to diagnose various diseases at an early stage, so that preventive action can be taken by individuals.
- The machine learning algorithms have showed high performance on predicting the diabetes [3, 8, 9, 10, 11], heart disease [12, 13, 14, 15], lung cancer [16, 17], and other diseases based on current conditions of individuals.

Disease	Author	Algorithm	Dataset	Accuracy (%)
Diabetes	Fitriyani et. al [3]	Forward Logistic Regression and MLP	NAGALA	92.11
	Patil et. al [8]	Decision Tree C4.5	The Pima Indians	92.38
	Wu et. al [9]	K-Mean and Logistic Regression	The Pima Indians	93.50
	Ijaz et. al [10]	DBSCAN+SMOTE+Random Forest	Dr John Schorling	92.55
	Fitriyani et. al [11]	iForest+SMOTETomek+Ensemble Learning	Dr John Schorling	100.00
Heart Disease	Bhatt et. al [12]	GridSearchCV+MLP	CVD (Kaggle)	87.28
	Fitriyani et. al [13]	DBSCAN+SMOTE-ENN+XGBOOST	Statlog Cleveland	95.90 98.40
	Ali et. al [14]	Stacked SVMs	Cleveland	92.22
	Gupta et. al [15]	FAMD + Random Forest	Cleveland	93.44
Lung Cancer	Dritsas and Trigka [16]	SMOTE + Rotation Forest	Lung Cancer (Kaggle)	97.10
	Alam et. al [17]	Watershed Transform + GLCM + SVM	Lung Cancer (UCI ML Rep)	97.00

02

Machine Learning

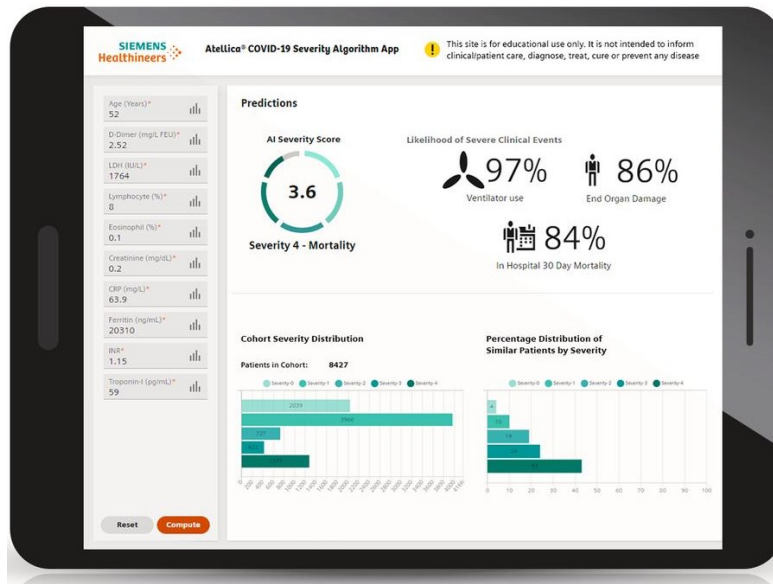
MS | LEARNING



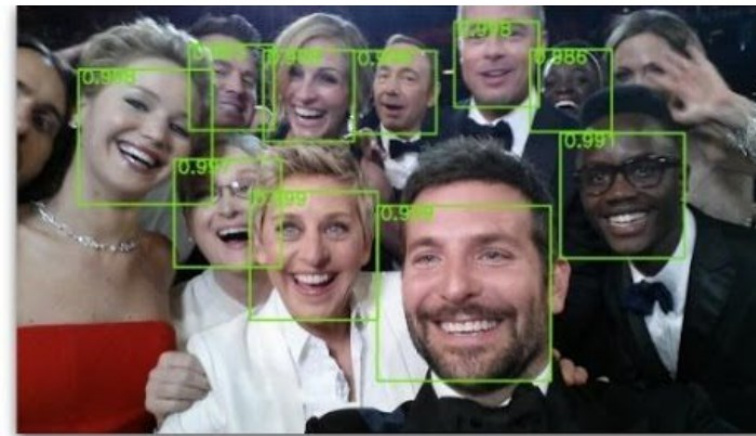
세종대학교
SEJONG UNIVERSITY

ITATS
INSTITUT
TEKNOLOGI
ADHI TAMA
SURABAYA

Machine Learning

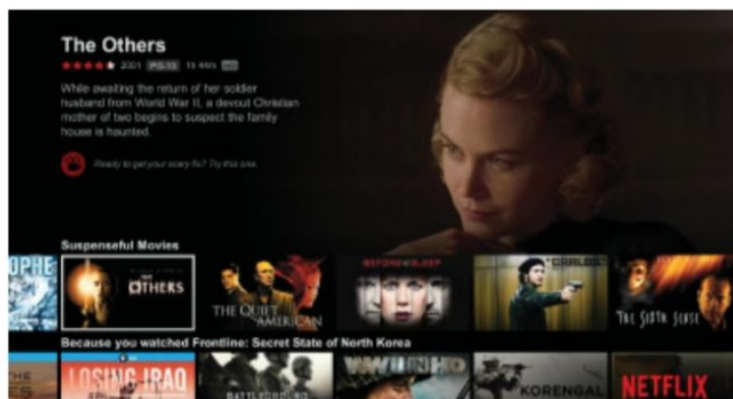


Machine learning is a powerful tool; it can...



Recognize your face in a photo.

Identify potential disease progression and predict disease



Recommend movies you will like.

Predict potential bank customer

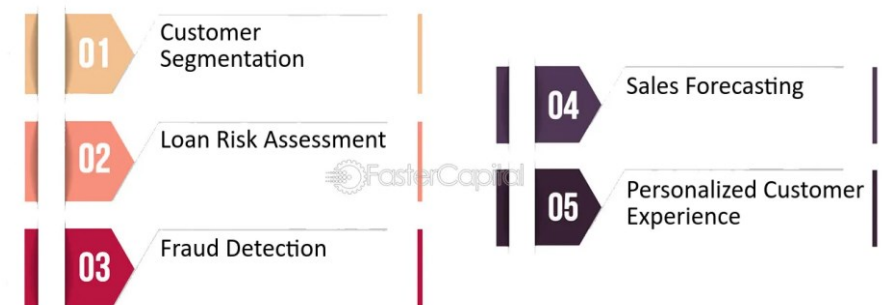
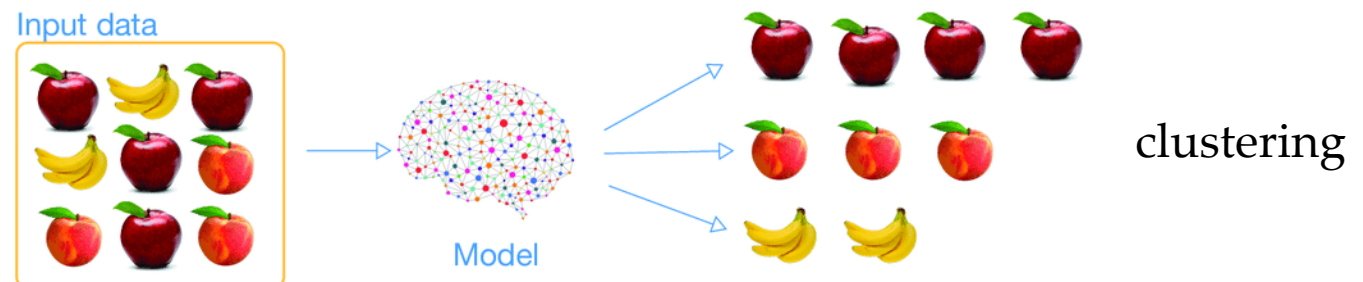
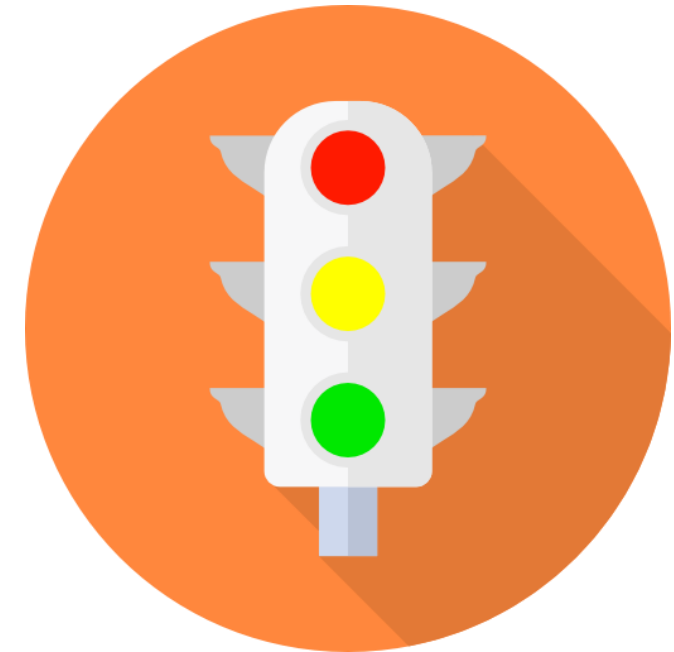
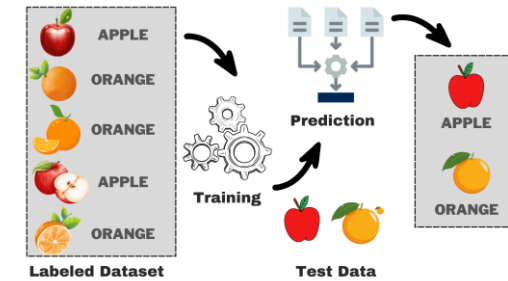
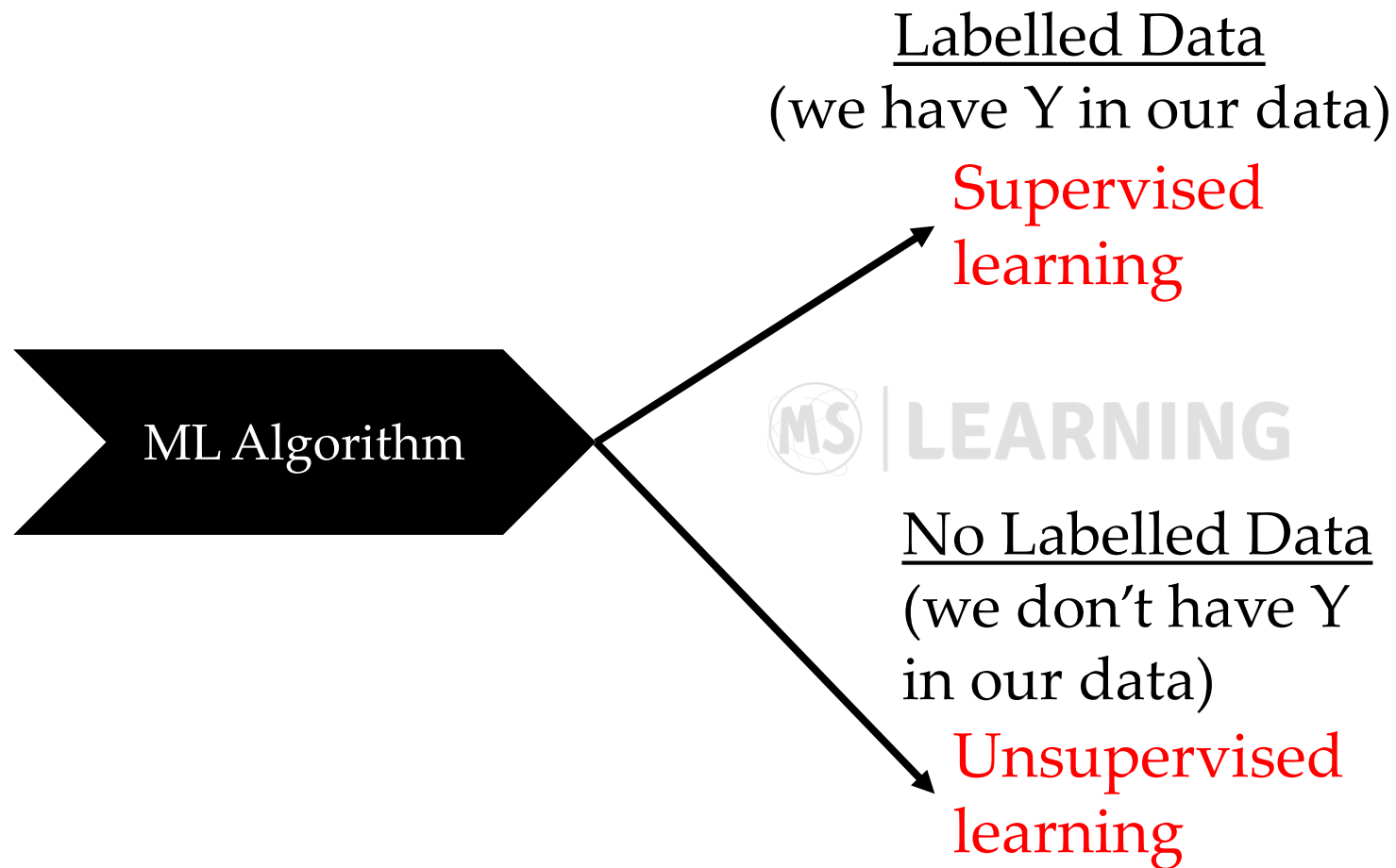


Fig 5 Real-case application of machine learning-based model in healthcare or medical

Machine Learning

Types of Machine Learning Algorithms



Machine Learning

Supervised learning task

The task function depends on the type of data the individual wants to predict. Supervised learning problems fall into two main categories: regression and classification.



The Task

Regression

Continuous variable

A regression problem is when we are trying to **predict a numerical value**, such as “stock price” or “blood sugar”

Classification

Categorical variable

A classification problem is when we are trying to **predict whether something belongs to a category**, such as “red” or “blue” or “disease” and “no disease”

↑
We are here!

Machine Learning

Supervised Machine Learning

- Logistic Regression
- Decision Tree
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Neural Network : Multilayer Perceptron (MLP)
- Naïve Bayes
- Random Forest
- AdaBoost
- Extreme Gradient Boosting (XGB)
- LightGBM
- CatBoost

MS | LEARNING

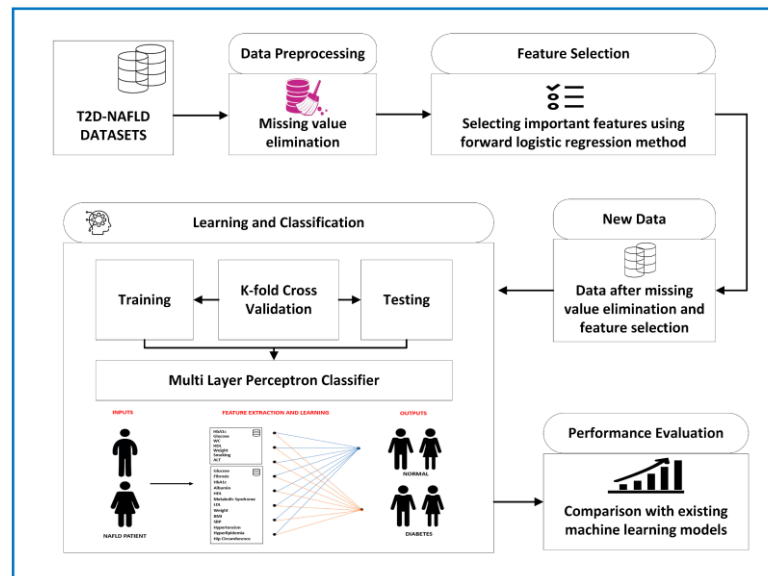
03

Machine Learning-based Disease Prediction Models

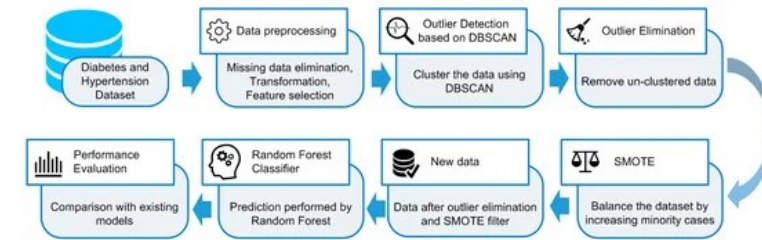


Development of machine learning-based disease prediction model

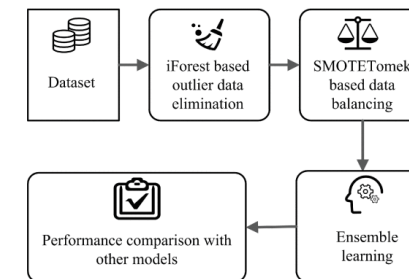
1. Data Collection
2. Data Preparation
3. Choose the Model/Algorithm
4. Training the Model
5. Evaluate the Model



(a) [3]



(b) [11]



(c) [12]



(d) [21]

Development of machine learning-based disease prediction model

1. Data Collection

There are two main categories of data [22]:

Primary Data

- Primary data is newly collected data;
- It can be gathered directly from people's responses (surveys), or from their biometrics (blood pressure, weight, blood tests, etc.).
- The data is collected for other (medical) purposes by extracting the data from medical records.

Secondary Data

- Secondary data is data that already exists;
- It has already been published or complied.
- There are extant local, regional, national and international databases such as Public Health Data, government statistics, and WHO data.
Public health data sources : UCI ML Repository, Kaggle, data.world, KHNES, NHIS, etc.

Development of machine learning-based disease prediction model

1. Data Collection

Public health dataset example:

Age	Gender	Body Mass Index	Obesity	Hypertension	Hyperlipidemia	Metabolic Syndrome	Smoking Status	AST	ALT	ALP	GGT	Total Cholesterol	Triglyceride	HDL	LDL	Glucose	Steatosis	Activity	Fibrosis	NAS score	NAS score	Fibrosis	Significance	Advanced	Cirrhosis	Diagnosis	Type of Disease (Mild illness)	Class
60	1	35.56	1	1	1	0	2	27	49	62	19	170	69	61	95	119	2	2	1	4	1	1	0	0	0	1	2	0
53	2	34.95	1	1	1	1	2	51	74	117	53	190	146	58	103	93	2	2	3	4	1	1	1	1	0	1	2	1
33	2	31.02	1	0	1	1	2	31	72	89	49	199	304	36	102	90	2	2	1	4	1	1	0	0	0	1	2	1
23	2	25.91	0	0	0	0	2	32	51	167	34	211	164	34	144	89	2	2	0	4	1	0	0	0	0	1	2	0
36	2	30.06	1	0	0	0	2	38	96	381	484	187	413	32	114	96	2	2	0	4	1	0	0	0	0	1	2	0
54	1	24.65	0	0	0	1	1	29	47	99	30	224	177	59	144	95	2	2	1	4	1	1	0	0	0	1	2	0
50	1	34.38	1	0	1	0	1	37	55	187	36	181	97	56	106	97	2	2	2	4	1	1	1	0	0	1	2	0
32	2	26.09	0	0	0	0	2	23	31	100	23	141	72	40	87	119	2	2	1	4	1	1	0	0	0	1	2	1
18	2	35.49	1	0	0	0	1	88	206	135	92	211	89	56	137	80	2	2	0	4	1	0	0	0	0	1	2	0
32	2	36.65	1	0	0	1	1	66	66	191	28	163	115	41	114	87	2	2	0	4	1	0	0	0	0	2	2	0
43	2	34.09	1	0	1	0	2	25	39	67	29	211	358	43	132	96	2	2	0	4	1	0	0	0	0	1	2	0
28	2	24.02	0	0	1	0	1	51	121	49	42	219	121	46	149	84	2	2	0	4	1	0	0	0	0	1	2	0
60	1	34.48	1	1	0	1	1	47	59	92	57	125	125	74	143	100	2	2	0	4	1	0	0	0	0	1	2	1
28	2	24.99	0	0	1	0	2	64	96	85	47	124	307	19	44	87	2	2	3	4	1	1	1	1	0	1	2	0
51	2	35.4	1	1	1	1	1	41	63	56	27	270	204	46	183	105	2	2	0	4	1	0	0	0	0	1	2	0
43	2	27.4	0	0	0	0	1	24	26	81	37	228	128	59	161	106	2	2	0	4	1	0	0	0	0	1	2	0
52	1	37.3	1	0	0	0	1	58	56	89	27	152	212	58	81	97	2	2	2	4	1	1	1	0	0	1	2	0
55	2	36.65	1	0	1	1	3	37	67	61	39	235	304	42	132	102	2	2	3	4	1	1	1	1	0	1	2	0
43	2	27.4	0	0	0	0	1	24	26	81	37	228	128	59	161	106	2	2	0	4	1	0	0	0	0	1	2	0
55	2	27.92	0	1	1	1	3	20	19	96	34	183	396	28	76	141	2	2	3	4	1	1	1	1	0	1	2	1
33	2	31.64	1	0	1	1	3	25	43	99	28	227	434	34	131	83	2	2	0	4	1	0	0	0	0	1	2	0
42	2	33.91	1	0	1	1	2	39	84	166	178	178	380	30	72	89	2	2	2	4	1	1	1	0	0	1	2	0
39	1	32.1	1	0	1	1	1	18	12	60	9	182	210	35	105	110	2	2	0	4	1	0	0	0	0	1	2	0
38	2	32	1	0	1	0	1	41	63	103	104	210	132	48	136	100	2	2	1	4	1	1	0	0	0	1	2	0
35	2	37.8	1	0	1	1	1	16	26	53	19	226	170	40	152	99	2	2	1	4	1	1	0	0	0	1	2	0
44	2	27.85	0	0	1	1	2	68	134	56	64	237	296	47	130	102	2	2	0	4	1	0	0	0	0	1	2	0
47	1	42.04	1	0	0	1	2	19	21	105	46	226	323	39	122	129	2	2	0	4	1	0	0	0	0	2	2	1

• • • • •

563	63	2	28.61	0	1	1	1	2	36	55	48	65	283	215	40	200	115	1	4	1	6	0	1	0	0	0	1	2	0
564	44	1	40.3	1	0	1	1	1	30	39	77	29	183	106	60	110	133	1	4	3	6	0	1	1	1	0	1	2	1
565	44	1	40.35	1	0	1	1	1	30	39	77	29	183	106	60	110	133	1	4	3	6	0	1	1	1	0	1	2	1
566	64	2	29.74	0	1	1	1	3	20	20	65	176	170	262	49	58	112	1	4	3	6	0	1	1	1	0	1	2	1

565 rows × 29 columns

Development of machine learning-based disease prediction model

2. Data Preparation

- Data preparation in machine learning is the process of cleaning, transforming, and organizing raw data into a format that machine learning algorithms can understand.
- Processing the data into good quality data [22] due to lack quality of data.

Missing value elimination

Microsoft Excel, Python, etc.

Duplicate data elimination

Microsoft Excel, Python, etc.

Noise or outlier elimination

Isolation Forest, DBSCAN, Local Outlier Factor, standard deviation, interquartile range

- Processing the data due to the high quantity of data (size or dimension) that could affect the performance of the model.

Imbalance class distribution
(high difference number of '+' and '-' class)

Too large data (number of features or dimension)

Over sampling Technique : SMOTE, ADSYN, Random Over Sampling (ROS)
Under sampling Technique : Tomek Link, Random Under Sampling, NearMiss
Hybrid Technique : SMOTE-ENN, SMOTE-Tomek

Feature selection : Information Gain, Chi Square Test, Fisher Score, Correlation Coefficient, Forward Selection, Backward Selection, Recursive Feature Elimination, Tree-based (RF, XGB)
Feature Extraction : Principal Component Analysis

Development of machine learning-based disease prediction model

3. Choose the Model/Algorithm

- In machine learning, choosing the right model is one of the most important steps in building a successful predictive model.
- Choosing the wrong model can lead to poor performance, wasted time and resources, and inaccurate results.



Development of machine learning-based disease prediction model

3. Choose the Model/Algorithm

- Steps to choose the right machine learning model:
 - **Define the problem** : the researcher needs to understand what kind of problem he/she is dealing with. Is it a classification problem or a regression problem? is he/she trying to predict a categorical or continuous outcome?
 - **Consider the data** : the researcher should know the *feature types* (numerical or categorical, text or image), different models may be better suited for different feature types. *Feature importance*: are all features equally important, or are some more important than others? If some features are more important, and want to use a model that can perform feature selection or feature weighting, such as random forests [23]. *Data size*: how much data does the researcher have? If the dataset is small, simpler models may be more appropriate to avoid overfitting [23]. If dataset is large, more complex models may be able to capture the patterns. *Data distribution*: Is the data distribution balanced or imbalanced?
 - **Evaluate different models or conducting model comparison**: each type of model has its own strengths and weaknesses, and it's important to evaluate each one carefully to determine which is best suited for the researcher's problem.

Development of machine learning-based disease prediction model

4. Training the Model

- Model training is the stage where the ML algorithm is trained by feeding datasets.
- Model training in machine language is the process of feeding an ML algorithm with data to help identify and demonstrate correlation between the input data and outcomes.

```
iteration 10
MLP
- Precision: 98.86363636363636 - Recall: 83.33333333333333 - F-1 Score: 89.42528735632185 - Accuracy: 97.82608695652173 - AUC: 0.8333333333333333 - sensitivity: 0.6666666666666666 - specificity: 1.0 - ppv: 1.0 - npv: 0.9772727272727273 - lrp: inf - lrm: 0.3333333333333337 - youden: 0.6666666666666665
LR
- Precision: 46.73913043478261 - Recall: 50.0 - F-1 Score: 48.31460674157303 - Accuracy: 93.47826086956522 - AUC: 0.5 - sensitivity: 0.0 - specificity: 1.0 - ppv: nan - npv: 0.9347826086956522 - lrp: nan - lrm: 1.0 - youden: 0.0
KNN
- Precision: 97.77777777777779 - Recall: 66.66666666666666 - F-1 Score: 73.86363636363636 - Accuracy: 95.65217391304348 - AUC: 0.6666666666666666 - sensitivity: 0.3333333333333333 - specificity: 1.0 - ppv: 1.0 - npv: 0.9555555555555556 - lrp: inf - lrm: 0.6666666666666667 - youden: 0.3333333333333326
DT
- Precision: 98.86363636363636 - Recall: 83.33333333333333 - F-1 Score: 89.42528735632185 - Accuracy: 97.82608695652173 - AUC: 0.8333333333333333 - sensitivity: 0.6666666666666666 - specificity: 1.0 - ppv: 1.0 - npv: 0.9772727272727273 - lrp: inf - lrm: 0.3333333333333337 - youden: 0.6666666666666665
SVM
- Precision: 97.77777777777779 - Recall: 66.66666666666666 - F-1 Score: 73.86363636363636 - Accuracy: 95.65217391304348 - AUC: 0.6666666666666666 - sensitivity: 0.3333333333333333 - specificity: 1.0 - ppv: 1.0 - npv: 0.9555555555555556 - lrp: inf - lrm: 0.6666666666666667 - youden: 0.3333333333333326
NB
- Precision: 98.86363636363636 - Recall: 83.33333333333333 - F-1 Score: 89.42528735632185 - Accuracy: 97.82608695652173 - AUC: 0.8333333333333333 - sensitivity: 0.6666666666666666 - specificity: 1.0 - ppv: 1.0 - npv: 0.9772727272727273 - lrp: inf - lrm: 0.3333333333333337 - youden: 0.6666666666666665
RF
- Precision: 98.86363636363636 - Recall: 83.33333333333333 - F-1 Score: 89.42528735632185 - Accuracy: 97.82608695652173 - AUC: 0.8333333333333333 - sensitivity: 0.6666666666666666 - specificity: 1.0 - ppv: 1.0 - npv: 0.9772727272727273 - lrp: inf - lrm: 0.3333333333333337 - youden: 0.6666666666666665
ADA
- Precision: 100.0 - Recall: 100.0 - F-1 Score: 100.0 - Accuracy: 100.0 - AUC: 1.0 - sensitivity: 1.0 - specificity: 1.0 - ppv: 1.0 - npv: 1.0 - lrp: inf - lrm: 0.0 - youden: 1.0
XGB
- Precision: 100.0 - Recall: 100.0 - F-1 Score: 100.0 - Accuracy: 100.0 - AUC: 1.0 - sensitivity: 1.0 - specificity: 1.0 - ppv: 1.0 - npv: 1.0 - lrp: inf - lrm: 0.0 - youden: 1.0
[LightGBM] [Info] Number of positive: 28, number of negative: 391
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000245 seconds.
You can set 'force_col_wise=true' to remove the overhead.
[LightGBM] [Info] Total Bins 520
[LightGBM] [Info] Number of data points in the train set: 419, number of used features: 10
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.066826 -> initscore=-2.636503
[LightGBM] [Info] Start training from score -2.636503
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
LGB
- Precision: 100.0 - Recall: 66.66666666666666 - F-1 Score: 80.0 - Accuracy: 97.82608695652173 - AUC: 0.8333333333333333 - sensitivity: 0.6666666666666666 - specificity: 1.0 - ppv: 1.0 - npv: 0.9772727272727273 - lrp: inf - lrm: 0.3333333333333337 - youden: 0.6666666666666665
```

Fig 13. Model training

Development of machine learning-based disease prediction model

5. Evaluating the Model

- The model evaluation is a step where the performance of the model is tested on previously unseen data.
- The unseen data used is the testing set that is split from the master dataset before model selection.
- The performance of the model is evaluated used numerous evaluation metrics in machine learning such as accuracy, precision or positive predictive value (ppv), recall or sensitivity or true positive rate (tpr), negative predictive value (npv), f1 score, specificity, area under the curve (AUC), etc.

MLP Final Acc: 96.559 Final Prec: 93.234 Final Spec: 74.583 Final Rec: 74.583 Final F1: 80.097 Final AUC: 0.746	LR Final Acc: 94.408 Final Prec: 67.185 Final Spec: 58.333 Final Rec: 58.333 Final F1: 60.048 Final AUC: 0.583	KNN Final Acc: 93.978 Final Prec: 61.975 Final Spec: 55.000 Final Rec: 55.000 Final F1: 55.938 Final AUC: 0.550	DT Final Acc: 96.785 Final Prec: 88.800 Final Spec: 88.247 Final Rec: 88.247 Final F1: 86.591 Final AUC: 0.882	SVM Final Acc: 96.341 Final Prec: 93.125 Final Spec: 72.917 Final Rec: 72.917 Final F1: 78.541 Final AUC: 0.729
NB Final Acc: 87.405 Final Prec: 83.009 Final Spec: 83.239 Final Rec: 83.239 Final F1: 78.524 Final AUC: 0.832	RF Final Acc: 96.989 Final Prec: 93.964 Final Spec: 83.704 Final Rec: 83.704 Final F1: 86.400 Final AUC: 0.837	ADA Final Acc: 96.998 Final Prec: 90.437 Final Spec: 86.810 Final Rec: 86.810 Final F1: 87.186 Final AUC: 0.868	XGB Final Acc: 97.428 Final Prec: 87.000 Final Spec: 90.144 Final Rec: 81.667 Final F1: 79.810 Final AUC: 0.901	LightGBM Final Acc: 96.984 Final Prec: 91.667 Final Spec: 82.154 Final Rec: 65.000 Final F1: 71.476 Final AUC: 0.822

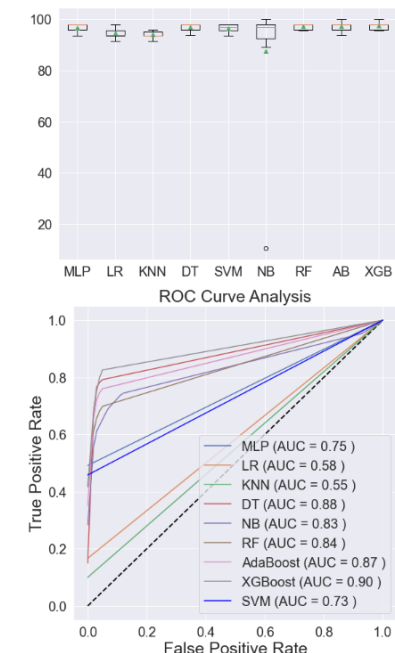


Fig 14. Model Performance

Development of machine learning-based disease prediction model

5. Evaluating the Model

Model	Performance evaluation							
	acc (%)	pre (%)	rec/sen/TPR (%)	f (%)	MCC	FPR (%)	FNR (%)	TNR (%)
NB	84.07 ± 4.70	84.36 ± 7.85	80.00 ± 9.28	81.61 ± 5.46	0.68 ± 0.10	12.67 ± 7.57	20.00 ± 9.28	87.33 ± 7.57
LR	84.81 ± 4.21	85.49 ± 7.38	80.83 ± 11.81	82.21 ± 5.71	0.70 ± 0.08	12.00 ± 7.18	19.17 ± 11.81	88.00 ± 7.18
MLP	85.56 ± 4.21	86.12 ± 6.52	81.67 ± 12.25	82.99 ± 6.03	0.67 ± 0.12	11.33 ± 6.00	18.33 ± 12.25	88.67 ± 6.00
SVM	69.63 ± 7.37	72.90 ± 11.29	50.83 ± 10.83	59.52 ± 10.17	0.38 ± 0.16	15.33 ± 7.33	49.17 ± 10.83	84.67 ± 7.33
DT	74.81 ± 8.57	74.28 ± 13.61	70.83 ± 12.50	71.39 ± 9.27	0.49 ± 0.17	3.33 ± 12.74	28.33 ± 11.90	76.67 ± 12.74
RF	82.96 ± 8.15	85.15 ± 10.71	75.83 ± 12.05	79.64 ± 9.70	0.68 ± 0.14	12.00 ± 8.33	23.33 ± 11.06	88.00 ± 8.33
Proposed HDPM	95.90 ± 5.55	97.14 ± 5.71	94.67 ± 11.08	95.35 ± 6.52	0.92 ± 0.10	4.52 ± 6.94	3.33 ± 6.67	95.48 ± 6.94

Note: acc = accuracy, pre = Precision, rec/sen/TPR = recall/ sensitivity/ true positive rate, f = f-measure, MCC = Matthews correlation coefficient, FPR = false positive rate, FNR = false negative rate, TNR = true negative rate.

(a)

Model	Performance evaluation							
	acc (%)	pre (%)	rec/sen/TPR (%)	f (%)	MCC	FPR (%)	FNR (%)	TNR (%)
NB	83.17 ± 7.64	84.18 ± 9.75	78.79 ± 8.29	81.25 ± 8.29	0.66 ± 0.15	13.12 ± 8.59	21.21 ± 8.29	86.88 ± 8.59
LR	84.85 ± 6.91	86.12 ± 7.85	80.22 ± 9.30	82.90 ± 7.80	0.70 ± 0.14	11.25 ± 6.73	19.78 ± 9.30	88.75 ± 6.73
MLP	84.15 ± 7.76	85.01 ± 9.74	80.22 ± 9.83	82.28 ± 8.66	0.68 ± 0.12	14.37 ± 9.29	18.41 ± 9.22	85.62 ± 9.29
SVM	71.06 ± 6.16	74.65 ± 9.51	59.23 ± 14.88	64.53 ± 9.43	0.43 ± 0.12	18.75 ± 10.46	40.77 ± 14.88	81.25 ± 10.46
DT	76.09 ± 4.86	74.21 ± 7.29	75.16 ± 8.00	74.31 ± 5.18	0.52 ± 0.09	24.38 ± 8.12	27.80 ± 7.31	75.62 ± 8.12
RF	82.14 ± 6.84	83.69 ± 8.63	76.54 ± 10.09	79.63 ± 8.36	0.66 ± 0.13	12.50 ± 8.39	22.03 ± 10.27	87.50 ± 8.39
Proposed HDPM	98.40 ± 3.21	98.57 ± 4.29	98.33 ± 5.00	98.32 ± 3.37	0.97 ± 0.06	1.67 ± 5.00	0.00 ± 0.00	98.33 ± 5.00

(b)

Fig 15. Performance of the machine learning models for predicting heart disease in combination with the Information Gain-based feature selection, DBSCAN-based outlier removal, SMOTE-ENN-based data balancing methods in Cleveland (a) and Statlog (b) datasets [13]

Classification model	Performance metric				
	p (%)	r (%)	f (%)	acc (%)	AUC
MLP	52.59	48.57	45.08	84.9	0.85
SVM	88.235	41.096	56.075	81.9	0.62
DT	36.37	42.68	32.18	69.69	0.59
LR	52.5	47.14	43.77	84.9	0.91
K-means + LR [16]	91.6	96.4	-	90.7	0.957
DBSCAN + SMOTE + RF [40]	91.497	93.403	92.440	92.555	-
Proposed DPM	94.49	98.62	96.32	96.74	0.99

(a)

Classification model	Performance metric				
	p (%)	r (%)	f (%)	acc (%)	AUC
MLP	73.23	97.56	83.56	72.02	0.5
SVM	73.04	99.17	84.1	72.58	0.45
DT	71.35	69.42	70.13	56.96	0.46
LR	73.67	95.06	82.77	71.33	0.58
CART [19]	-	58.38	-	-	0.68
DBSCAN + SMOTE + RF [40]	78.788	70.270	74.286	76.419	-
Proposed DPM	93.57	84.89	88.8	85.73	0.87

(b)

Classification model	Performance metric				
	p (%)	r (%)	f (%)	acc (%)	AUC
MLP	57.31	69.81	57	54.92	0.53
SVM	56.38	85.26	67.77	53.48	0.44
DT	57.33	54.1	54.69	49.96	0.49
LR	61.53	82.76	69.52	59.41	0.62
CART [19]	-	45.65	-	-	0.566
Proposed DPM	75.6	81.78	77.12	75.78	0.76

(c)

Classification model	Performance metric				
	p (%)	r (%)	f (%)	acc (%)	AUC
MLP	67.78	70.49	67.02	80.84	0.89
SVM	64.49	57.69	56.66	75.85	0.77
DT	61.48	56.43	56.13	72.48	0.75
LR	67.98	77.14	69.67	81.57	0.89
DBSCAN + SMOTE + RF [40]	83.665	84.677	84.168	83.644	-
Proposed DPM	100	100	100	100	1

(d)

Fig 16. Performance of the machine learning models for predicting T2D and hypertension in combination with the Information Gain-based feature selection, iForest-based outlier removal, SMOTETomek-based data balancing methods [11] in Dr John Schorling (a), Golino et al male hypertension (b), Golino et al female prehypertension (c), and Dr. P. Soundarapandian, M.D., D.M CKD datasets [11]

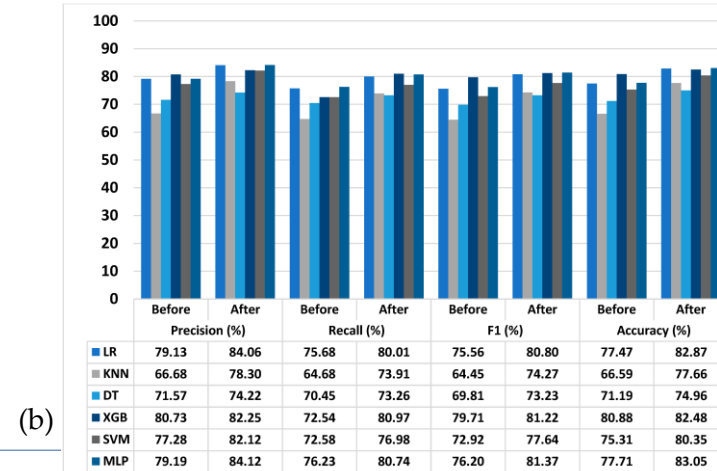
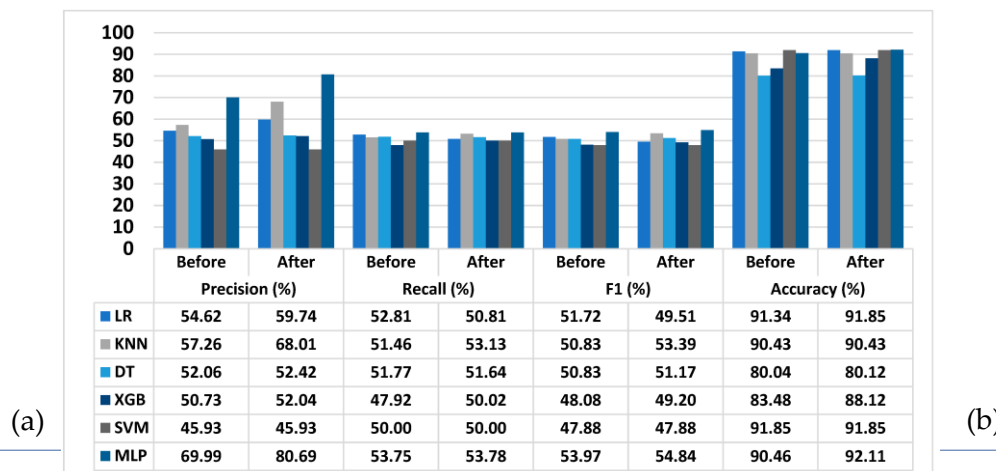


Fig 17. Performance of the machine learning models for predicting T2D in patient with NAFLD in combination with forward logistic regression in NAGALA (a) and NAFLD (b) datasets [3]

Important

What should be concerned when evaluating the disease prediction model's performance?

- In order to evaluate the classification model's performance, a summarized table called the **confusion matrix** is used.
- The confusion matrix consists of four categories:
 - True Negative (TN)** represents the number of samples correctly classified or predicted as belonging to the negative class. For example, the actual class is negative (0), and the predicted class is also negative (0).
 - True Positive (TP)** represents the number of samples correctly classified or predicted as belonging to the positive class. For example, the actual class is positive (1), and the predicted class is also positive (1).
 - False Negative (FN)** represents the number of samples incorrectly predicted as the negative class. For example, the actual class is positive (1), but the predicted class is negative (0).
 - False Positive (FP)** represents the number of samples incorrectly predicted as the positive class. For example, the actual class is negative (0), but the predicted class is positive (1).
- According to Hicks et al [24], the most commonly used for evaluating the performance of the ML-based disease prediction model are accuracy, recall or sensitivity or true positive rate (tpr), precision or positive predictive value (ppv), negative predictive value (npv), f1 score, Matthew's correlation coefficient (MCC), and threat score (TS).

MODEL'S PREDICTED VALUE			
		0	1
ACTUAL VALUE	0	True negative (TN)	False positive (FP)
	1	False negative (FN)	True positive (TP)

Performance Evaluation

confusion matrix

y6 (predict) = y'	y target (actual) = y
1	1
1	0
1	1
1	0
1	0
1	0
1	0
1	1
1	0
1	1

		MODEL'S PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	True negative (TN) 0	False positive (FP) 1
	1	False negative (FN) 0	True positive (TP) 1

- **Accuracy** : percentage of correctly classified samples over the total number of samples. Accuracy measures the overall correctness of the model's predictions.

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN}$$

$$\text{Testing Accuracy} = \frac{0+1}{0+1+1+0} = 0.5 = 50 \%$$

- **Recall/Sensitivity/TPR** : the ratio between correctly classified positive samples and all samples assigned to the positive class [23]. When it's actually yes, how often does it predict yes?

$$\text{Recall} = \frac{TP}{\text{Actual Yes (1)}} = \frac{TP}{TP+FN}$$

$$\text{Testing Recall} = \frac{1}{1+0} = 1 = 100 \%$$

Performance Evaluation

confusion matrix

y6 (predict) = y'	y target (actual) = y
1	1
1	0
1	1
1	0
1	0
1	0
1	0
1	1
1	0
1	1

		MODEL'S PREDICTED VALUE	
		0	1
ACTUAL VALUE	0	True negative (TN) 0	False positive (FP) 1
	1	False negative (FN) 0	True positive (TP) 1

- **Precision/Positive Predictive Value (PPV)** : the ratio between correctly classified samples and all samples assigned to that class. When it predicts yes, how often is it correct?

$$\text{Precision} = \frac{TP}{\text{Predicted Yes (1)}} = \frac{TP}{TP+FP}$$

$$\text{Testing Precision} = \frac{1}{1+1} = 0.5 = 50 \%$$

- **Negative Predictive Value (NPV)** : the ratio between correctly classified negative samples and all samples classified as negative. When it predicts no, how often is it correct?

$$\text{NVP} = \frac{TN}{\text{Predicted No(0)}} = \frac{TN}{TN+FN}$$

$$\text{Testing NPV} = \frac{0}{0+0} = \frac{0}{0} = 0 = 0 \%$$

Performance Evaluation

confusion matrix

y6 (predict) = y'	y target (actual) = y
1	1
1	0
1	1
1	0
1	0
1	0
1	0
1	1
1	0
1	1

MODEL'S PREDICTED VALUE			
		0	1
ACTUAL VALUE	0	True negative (TN) 0	False positive (FP) 1
	1	False negative (FN) 0	True positive (TP) 1

- **F1 score** : represents the harmonic mean or weighted average of precision and recall. A large F1 score of 1 indicates excellent precision and recall, while a low score indicates poor model performance.

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Testing F1} = \frac{2 \times 0.5 \times 1}{0.5 + 1} = \frac{1}{1.5} = 0.6667 = 66.67 \%$$

- **Specificity/True Negative Rate** : how often the model predicts a negative for a value that is actually negative.

$$\text{Specificity} = \frac{TN}{\text{Actual No}(0)} = \frac{TN}{TN + FP}$$

$$\text{Testing Specificity} = \frac{0}{0 + 1} = 0 = 0 \%$$

04

Machine Learning-based Model for Disease Prediction Applications



Machine Learning-based Model for Disease Prediction Applications

Heart Disease Clinical Decision Support System - Diagnosis Form

1. Patient (id): 200313
 2. Age (age): 49
 3. Gender (sex): ☒ Female (0) ☐ Male (1)
 4. Chest pain type (cp):
☐ Typical angina (1)
☐ Atypical angina (2)
☒ Non-anginal pain (3)
☐ Asymptomatic (4)
 5. Resting electrocardiographic results (restecg):
☐ Normal (0)
☒ Having ST-T wave abnormality (≥ 0.05 mV) (1)
☐ Showing probable or definite left ventricular hypertrophy by Estes' criteria (2)
 6. Maximum heart rate (tmax): 125
 7. Exercise induced angina (exang):
☐ No (0) ☒ Yes (1)
 8. ST depression induced by exercise relative to rest (oldpeak): 2
 9. The slope of the peak exercise ST segment (slope): 4.5
 10. Number of major vessels (0-3) colored by fluoroscopy (ca):
☐ 0 ☐ 1 ☒ 2 ☐ 3
 11. Defect type (thal):
☐ Normal (3)
☒ Fixed defect (6)
☐ Reversible defect (7)
 DIAGNOSE

Heart Disease Clinical Decision Support System - Diagnosis Result

Diagnosis Data

1. Patient (id): 200313
 2. Age (age): 49
 3. Gender (sex): Female (0)
 4. Chest pain type (cp): Non-anginal pain (3)
 5. Resting electrocardiographic results (restecg): Having ST-T wave abnormality (≥ 0.05 mV) (1)
 6. Maximum heart rate (tmax): 125
 7. Exercise induced angina (exang): Yes (1)
 8. ST depression induced by exercise relative to rest (oldpeak): 2
 9. The slope of the peak exercise ST segment (slope): 4.5
 10. Number of major vessels (0-3) colored by fluoroscopy (ca): 2
 11. Defect type (thal): Fixed defect (6)
Diagnosis Result
 Heart disease: **PRESENCE**

Fig 20. Web-based heart disease clinical support system [3]

Diseases prediction app

ID: PA101
 Date of birth: Jan 12, 1990
 Height (cm): 160
 Waist circumference (cm): 84
 Hip circumference (cm): 93
 Get weight data
 BLE weight sensor: CA:23:CA:A3:...
 Weight (kg): 54
 Send

Information

- ID : PA101
- Age : 29
- BMI : 21.1
- Is obese? : no
- Waist hip ratio : 0.90

Prediction result

- Hypertension : no
- Diabetes : no

Suggestion

Keep continue **eating healthy diet food** and do **regular exercise**. Regular exercise could help prevent diseases such as heart disease, stroke, diabetes, etc.

Fig 19. Web-based disease prediction application [11]

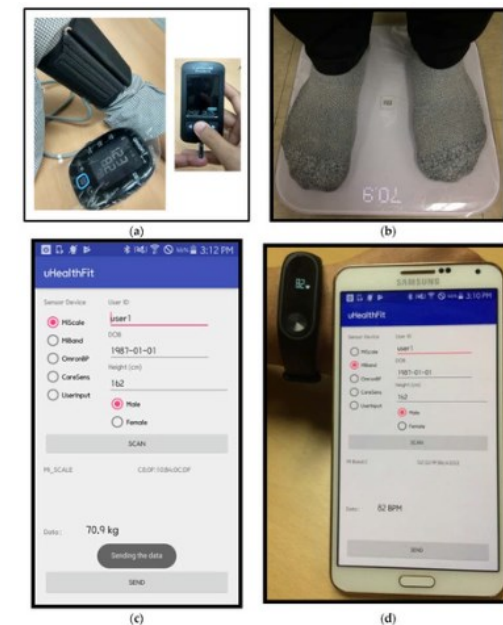
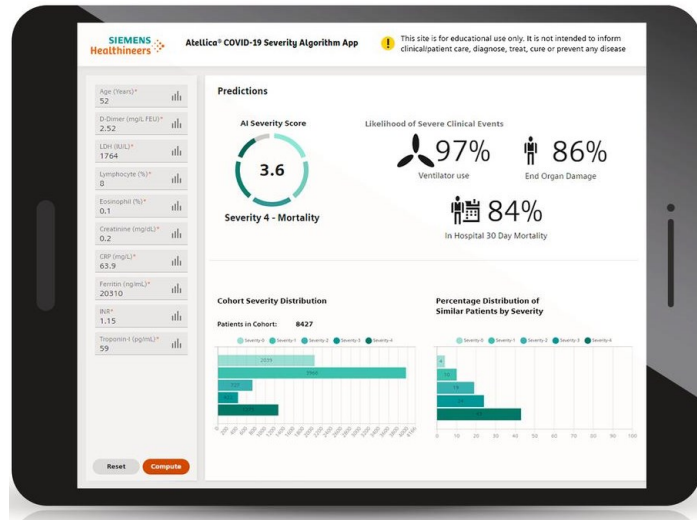
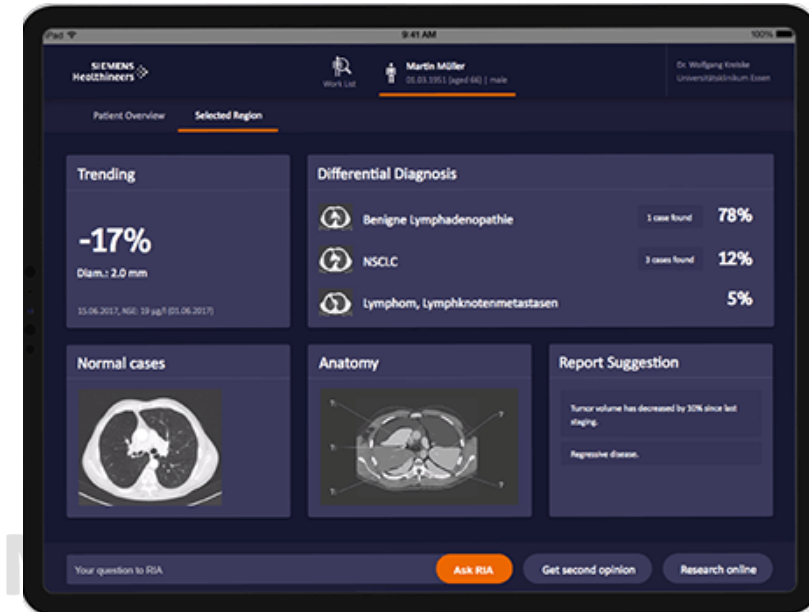


Fig 21. Personalize healthcare monitoring system [25]

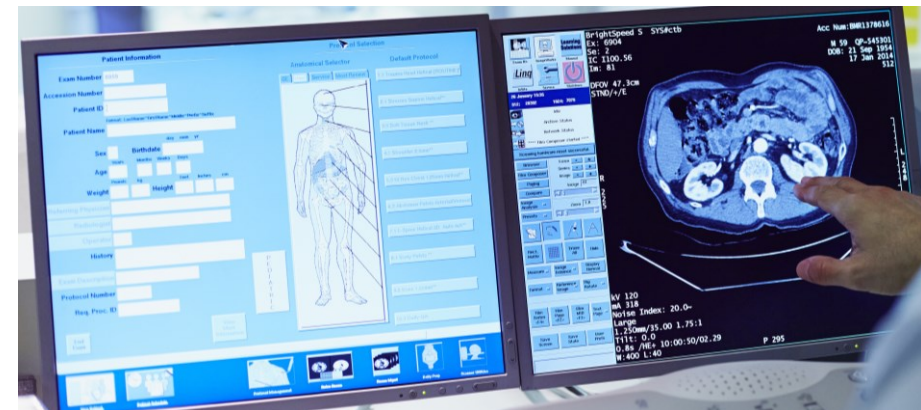
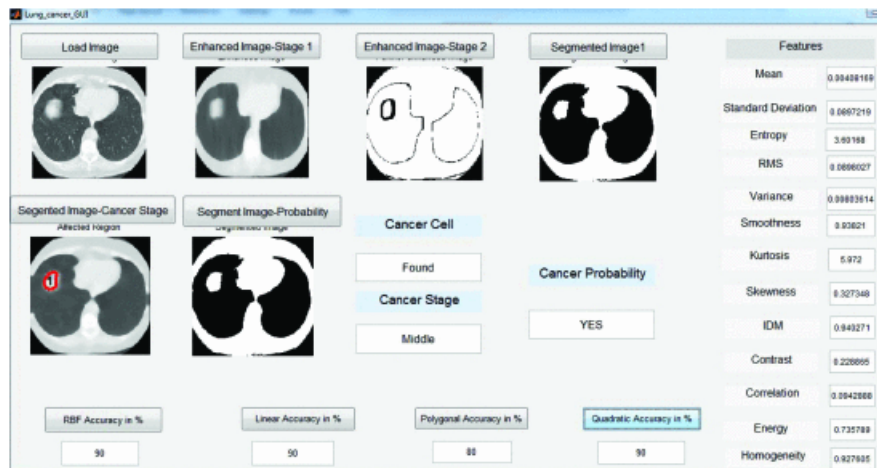
Machine Learning-based Model for Disease Prediction Applications



(a) Siemens Healthineers - Prediction and early identification of disease: identify potential disease progression in COVID-19 patient [18]



(b) Siemens Healthineers - Prediction and early identification of lung disease [19]



(c) and (d) Health system – detection and prediction lung cancer utilized by medical institutions [17, 20]

Fig 22. Real-case application of machine learning-based model in healthcare or medical

05

Conclusion



05 Conclusion



- Machine learning is a powerful tool that can be utilized as one of the alternatives to early detection of the disease.
- By utilizing ML as a disease prediction tool, it could help individuals know their current health status, thus preventing the occurrence of the worst-case scenario.
- Not only in the healthcare or medical domain, but machine learning has also been widely utilized in many other domains.

06 References



1. The top 10 causes of death. Available online: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. (Accessed on 23 April 2024)
2. Fitriyani, N.L.; Syafrudin, M.; Ulyah, S.M.; Alfian, G.; Qolbiyani, S.L.; Anshari, M. A Comprehensive Analysis of Chinese, Japanese, Korean, US-PIMA Indian, and Trinidadian Screening Scores for Diabetes Risk Assessment and Prediction. *Mathematics* 2022, 10, 4027. <https://doi.org/10.3390/math10214027>
3. Fitriyani, N.L.; Syafrudin, M.; Ulyah, S.M.; Alfian, G.; Qolbiyani, S.L.; Yang, C.-K.; Rhee, J.; Anshari, M. Performance Analysis and Assessment of Type 2 Diabetes Screening Scores in Patients with Non-Alcoholic Fatty Liver Disease. *Mathematics* 2023, 11, 2266. <https://doi.org/10.3390/math11102266>
4. Fendi, Ferry & Kurniati, Anna. (2021). Human Resources for Health Country Profiles of Indonesia.
5. Current Health Expenditure. Available online: <https://www.mohw.go.kr/menu.es?mid=a20311000000>. (Accessed on 23 April 2024)
6. Overview of the Medical Expense Insurance Market in Korea. Available online: https://eng.koreanre.co.kr/sub.asp?maincode=501&sub_sequence=519&sub_sub_sequence=575&exec=view&strBoardID=kui_575&intCategory=0&strSearchCategory=|s_name|s_subject|&strSearchWord=&intPage=1&intSeq=1680. (Accessed on 23 April 2024)
7. Massive Growth in Expenses and Rising Inflation Fuel Continued Financial Challenges for America's Hospitals and Health Systems. Available online: <https://www.aha.org/guidesreports/2023-04-20-2022-costs-caring>. (Accessed on 23 April 2024)
8. Patil, B.M.; Joshi, R.C.; Toshniwal, D. Hybrid prediction model for Type-2 diabetic patients. *Expert Syst. Appl.* 2010, 37, 8102–8108. [Google Scholar] [CrossRef]
9. Wu, H.; Yang, S.; Huang, Z.; He, J.; Wang, X. Type 2 diabetes mellitus prediction model based on data mining. *Inform. Med. Unlocked* 2018, 10, 100–107. [Google Scholar] [CrossRef]
10. Ijaz, M.F.; Alfian, G.; Syafrudin, M.; Rhee, J. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. *Appl. Sci.* 2018, 8, 1325. <https://doi.org/10.3390/app8081325>.
11. Fitriyani, N.L.; Syafrudin, M.; Alfian, G.; Rhee, J. Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access* 2019, 7, 144777–144789.
12. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* 2023, 16, 88. <https://doi.org/10.3390/a16020088>

06 References

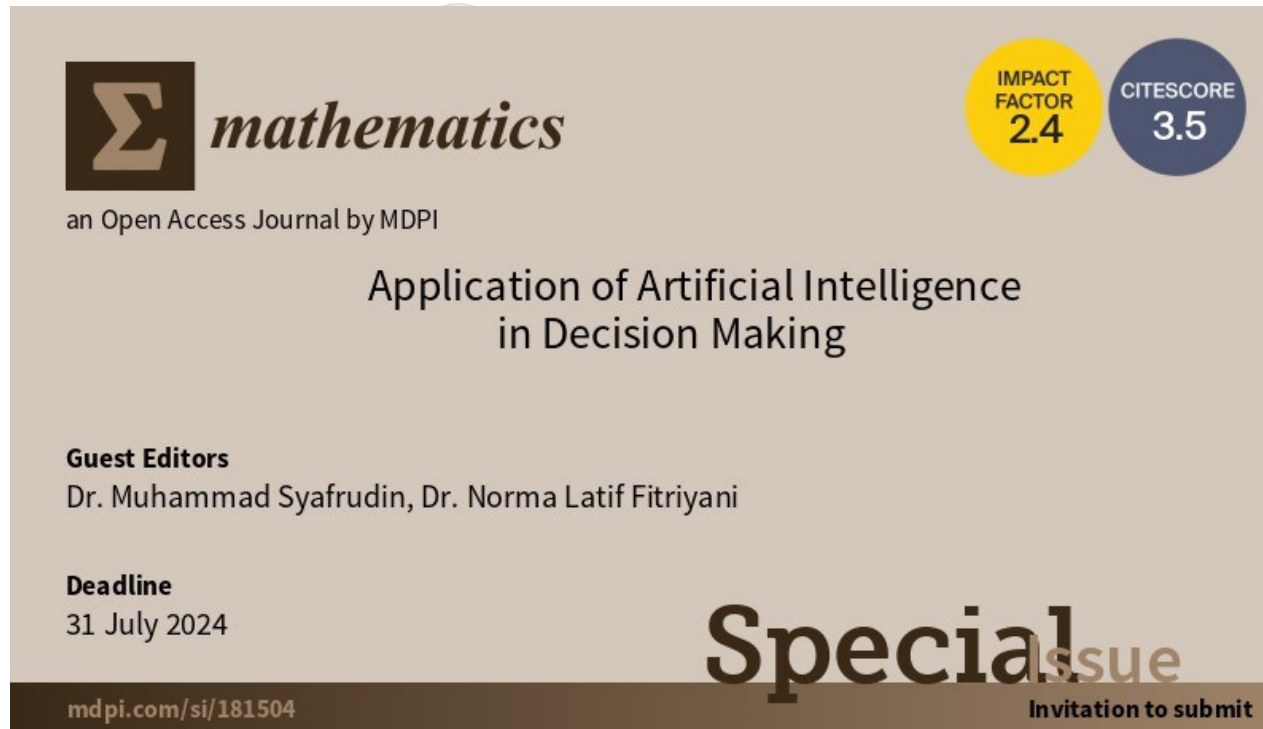


13. N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: an effective heart disease prediction model for a clinical decision support system," IEEE Access, vol. 8, Article ID 133034, 2020.
14. L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, et al., "An optimized stacked support vector machine s based expert system for the effective prediction of heart failure", IEEE Access, vol. 7, pp. 54007-54014, 2019.
15. A. Gupta, R. Kumar, H. S. Arora and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis", I IEEE Access, vol. 8, pp. 14659-14674, 2020.
16. Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. *Big Data Cogn. Comput.* **2022**, *6*, 139. <https://doi.org/10.3390/bdcc6040139>
17. J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465593.
18. Prediction and early identification of disease through artificial intelligence (AI). <https://www.siemens-healthineers.com/bg/digital-health-solutions/artificial-intelligence-in-healthcare/ai-to-help-predict-disease>. (Accessed on 23 April 2024)
19. Case Study Siemens Healthineers. <https://www.usability.de/en/clients/casestudies/case-study-siemens-healthineers-ux-design.html>. (Accessed on 23 April 2024)
20. https://storkclips.net/product_details/25498158.html. (Accessed on 23 April 2024)
21. Ghobadi, Fatemeh & Kang, Doosun. (2023). Application of Machine Learning in Water Resources Management: A Systematic Literature Review. *Water*. 15. 620. 10.3390/w15040620.
22. Vicken, T., Erin LS., Mohammad J., Hendry R. S. 2020. Acquiring data in medical research: A research primer for low- and middle-income countries. *African Journal of Emergency Medicine*, Vol.10, Pp. S135-S139. <https://doi.org/10.1016/j.afjem.2020.09.009>.
23. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: A synthetic knowledge synthesis. *Sci Prog.* 2022 Jan-Mar;105(1):368504211029777. doi: 10.1177/00368504211029777. PMID: 35220816; PMCID: PMC10358596.
24. Hicks, S.A., Strümke, I., Thambawita, V. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 12, 5979 (2022). <https://doi.org/10.1038/s41598-022-09954-8>
25. Alfian, G.; Syafrudin, M.; Ijaz, M.F.; Syaekhoni, M.A.; Fitriyani, N.L.; Rhee, J. A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing. *Sensors* 2018, 18, 2183. <https://doi.org/10.3390/s18072183>

Thank you for your attention!

Any questions?

Feel free to send any research/project collaboration proposals via norma@sejong.ac.kr



Σ mathematics
an Open Access Journal by MDPI

IMPACT FACTOR 2.4 CITESCORE 3.5

Application of Artificial Intelligence
in Decision Making

Guest Editors
Dr. Muhammad Syafrudin, Dr. Norma Latif Fitriyani

Deadline
31 July 2024

Special Issue
Invitation to submit

mdpi.com/si/181504