Anchor University Journal of Science and Technology (AUJST)
A publication of the Faculty of Science and Science Education, Anchor University Lagos

URL: fnas.aul.edu.ng

In AJOL: https://www.ajol.info/index.php/aujst

Vol. 4 No 1, September 2023, Pp. 52 - 70

ISSN: 2736-0059 (Print); 2736-0067 (Online)

AN IMPROVED DATA PRIVACY AND DATA AVAILABILITY MODEL FOR MEDICAL DIAGNOSIS SYSTEM USING A HYBRID WEIGHTED KNN AND RULE-BASED ALGORITHM

[1]UMAR, B. Umar, [1]Dutse, A. Yusuf and [2]Noma, M. Adamu

[1]Abubakar Tafawa Balewa University,

[2]Bauchi State University Gadau,

***Corresponding author:**
adamnoma@basug.edu.ng

Submitted 08 June, 2023
Accepted 21 July, 2023

Competing Interests:
The authors declare no competing interests.

ABSTRACT

Background: This study focuses on improving medical diagnosis systems, with a particular emphasis on addressing the challenges of data availability and data privacy associated with medical systems.
Objective: The goal is to develop a model that can be trained on large amounts of data and can accurately diagnose medical conditions while ensuring the privacy of patient data.
Method: To achieve this objective, we employ a combination of techniques, including the use of synthetic data generated of 100,000 samples from a sample of 4,390 by the Synthpop package.
Results: The synthetic data closely mimics the characteristics of the original observations, enabling us to overcome the limitations of limited data availability. This allows researchers to perform analysis without directly accessing sensitive patient information. Additionally, this research introduces an approach to protect patient privacy in clinical data sharing. It explores techniques for encapsulating data that maintains the statistical properties of the original data, allowing researchers to perform analysis without directly accessing sensitive patient information.
Conclusions: The hybrid weighted KNN with Rule-based model outperforms other conventional models by achieving an accuracy of 98% on the training data and 98% on the test data, 100% on precision and recall.

Keywords: Medical-diagnosis, medical-systems, patient, Synthpop package

1. INTRODUCTION

Disease diagnosis and treatment are essential tasks for medical consultants. However, errors in diagnosis can lead to incorrect drug prescriptions and complications in patient health. Furthermore, significant time is often spent on physical examinations and patient interviews before treatment can begin. To improve the accuracy and efficiency of medical diagnosis, various techniques have been developed, including hybrid medical expert systems that utilize both artificial and non-artificial approaches (Imhanlahimi & Otumu, 2019). Expert Systems have proven to be highly valuable in medical diagnosis (Nnebe *et al*., 2019). While traditional classification algorithms such as logistic regression and decision trees have been used for training medical diagnosis systems (Aswal *et al.,* 2016). Researchers are increasingly exploring the application of machine learning algorithms to enhance classification accuracy. However, the limited size of medical datasets poses challenges for training these medical systems (Collins *et al.,* 2017). Machine learning classifiers require a larger amount of data to

https://dx.doi.org/10.4314/aujst.v4i1.7

train on. (Torgyn & Khovanova, 2017).

This study aims to address three key issues related to medical diagnosis systems. Firstly, it will focus on mitigating the problem of limited data availability thereby curbing the issue of overfitting which arises when models are trained on small datasets. By addressing this challenge, the accuracy of the system can be improved. Secondly, this study also aims to propose a mechanism for addressing the issue of data privacy in the Medical Diagnosis System. As medical data contains sensitive and personal information, ensuring privacy and security is of utmost importance. By addressing the data privacy concern, the proposed medical diagnosis system can maintain the confidentiality and integrity of patient information, while still providing accurate and efficient diagnoses. Finally, the study proposes the design of an improved medical system using a hybrid weighted KNN and rule-based algorithm.

1.1 Data privacy challenges Associated with Medical Datasets

Data privacy is a critical concern in the context of Medical Diagnosis Systems, where sensitive patient information needs to be protected. The limited size of medical data sets can be attributed to the absence of comprehensive medical databases, as noted by Collins et al. (2017). Patient data is maintained and safeguarded by hospitals, insurance providers, clinics, and research institutions, resulting in fragmentation that is primarily driven by privacy legislation such as HIPAA (2013). These small-data sets typically comprise fewer than 500 observations, and in some cases as few as 10. Model training for machine learning algorithms such as ANN involves the machine's experience and exposure to data observations, with the goal of enabling the machine to learn. Subsequently, models are tested using new data to measure their performance and ensure they can be generalized to novel datasets (Collins et al., 2017). Several research articles have focused on addressing data privacy issues and proposing solutions for preserving patient privacy in medical diagnosis systems

Ateniese et al (2020) in their work proposed a privacy-preserving framework that utilizes techniques such as homomorphic encryption and secure multi-party computation. These methods enable collaborative analysis of medical data while preserving privacy. Qiu et al., (2018) privacy preservation in medical diagnosis is achieved through the use of distributed computing techniques. The article presents a privacy-preserving algorithm that enables collaborative diagnosis without revealing sensitive patient information to all parties involved.

Chen et al. (2023) focuses on privacy preservation in healthcare data through the application of techniques like secure aggregation and differential privacy. These privacy-preserving machine learning methods are explored for their applicability in medical diagnosis systems. Additionally, Tao et al (2010) addresses secure and privacy-preserving medical data sharing and analysis in a cloud computing environment. The article proposes cryptographic techniques such as homomorphic encryption and secure multi-party computation

https://dx.doi.org/10.4314/aujst.v4i1.7

to enable collaborative analysis while protecting patient privacy.

1.2 Data availability challenges associated with medical diagnosis systems

Das et al. (2020) tried to address the problem of overfitting in the medical diagnosis system using Neuro-fuzzy with a feature extraction model for classification. However, the system outperforms other models. The problem is that during feature extraction, there is a possibility of losing important information present in the original data and Feature extraction techniques may not scale well with increasing dataset sizes.in our work, we propose the use of feature selection. Sabay *et al* (2018) also tried to address the issue of overfitting in training medical diagnosis systems by Using Surrogate Data. Although similar to our work, this paper only focused on heart disease by augmenting the features of the Cleveland dataset. It also did not address the data privacy issue in medical diagnosis systems.

Hastie *et al.* (2009) introduced the L1 and L2 regularization methods, demonstrating their effectiveness in reducing overfitting in medical diagnosis tasks. Srivastava *et al.* (2014) proposed dropout regularization, a technique that randomly drops units during training, effectively preventing overfitting in deep neural networks. However, Excessive use of regularization techniques, such as L1 and L2 regularization or dropout, leads to underfitting, where the model fails to capture important patterns in the data. Perez and Wang (2017) explored the use of data augmentation techniques, such as image rotations, translations, and noise addition, to improve the generalization performance of medical image diagnosis systems. However, Data augmentation relies on the assumption that artificially generated data will adequately represent the underlying distribution of the problem. However, if the augmented data does not reflect the true variation in the target population, it may introduce bias or produce unrealistic samples.

Bergstra and Bengio (2012) presented the concept of hyperparameter optimization, including grid search and Bayesian optimization, to select optimal models and hyperparameter settings for medical diagnosis tasks. Model selection and hyperparameter tuning techniques, such as cross-validation and grid search, can be computationally expensive, particularly when dealing with large-scale medical datasets and complex models. It may require significant computational resources and time to identify the optimal configuration. Raghu *et al.* (2019) investigated the application of transfer learning in medical diagnosis systems, demonstrating the benefits of pre-training models on large datasets, such as ImageNet, and fine-tuning them for specific medical tasks. Transfer learning relies on the assumption that pre-trained models can effectively transfer knowledge to the target medical diagnosis task. However, the domain shift between the pre-training dataset (ImageNet) and the medical domain may limit the effectiveness of transferred features. Careful selection and fine-tuning of the pre-trained model are crucial to avoid negative transfer and ensure optimal performance.

In their 2016 publication entitled "Handling Limited Datasets with Neural Networks in

https://dx.doi.org/10.4314/aujst.v4i1.7

Medical Applications: A Small-Data Approach", Torgyn and Khovanova developed a novel solution to address the challenge of data volume requirements for machine learning models. The authors established a framework for generating surrogate data from small data sets, with a minimum of ten observations, which was validated using neural network techniques. This approach involves multiple runs of 2000 neural networks to produce robust data sets that replicate the characteristics of the original data set, thereby providing sufficient data volumes for modern machine learning-based predictions. However, the computational resources required for this technique, including the number of neurons, are substantial. Therefore, alternative solutions for surrogate data generation were explored, given the availability of various tools such as Synthpop, an R language library that offers data synthesis and comparison functions (Nowok et al., 2016).

In all of the aforementioned investigations, the primary focus of the researchers was predominantly on the machine learning algorithms employed in forecasting the ailment. The datasets utilized by these models were not of sufficient magnitude to facilitate training on Neural networks. The studies in question seemed to disregard methods that could be implemented on the dataset utilized by the models to enhance the precision and dependability of the prognostic accuracy. Therefore, our model proposes to provide a holistic solution to both the data privacy and data availability issue.

2. METHODOLOGY

This study experimented with the proposed model on Windows 10 Pro operating system. The OS runs on a personal computer (PC) of Intel Core i3 processor, with a speed of 2.30GHz. The proposed model was built from Python Scikit-learn machine learning libraries using Python 3. Jupyter Notebook 3 was used as the computational environment for this study. The hosting machine was a computer on Windows 10 Enterprise with Intel® Pentium(R) Dual-Core T450 @ CPU 2.30 GHz and 8 GB of RAM. All algorithms were implemented based on the adopted Scikit-learn library, including NumPy, SciPy, Scikit-learn, pandas, Matplotlib, Synthpop libraries.

2.1 Dataset definition

A publicly accessible dataset from Kaggle was used in the experiment. The collection had about 1000 distinct symptoms and over 230 different illnesses. The symptoms, age, and gender of a person served as the input for the machine learning algorithms. The symptoms, age, and gender of an individual were used as input to the machine learning algorithms.

2.1.2 Data processing

Firstly, to train our model, we clean the dataset. Originally, the dataset contained 132 symptoms (features) and 41 different types of diseases (target variable) for prediction. Next, we took a look at all the various symptoms and their prognosis to see which symptoms are very telling in sign. The distribution is shown in Table 1. From Table 1, shows Fatigue and vomiting are the two most common and most generic symptoms in this dataset and probably won't be a unique significant predictor for an

https://dx.doi.org/10.4314/aujst.v4i1.7

Table 1: Frequency of Symptoms in the dataset

No of sample	Symptom	Prognosis	Length
41	Fatigue	[Diabetes, Bronchial Asthma, Jaundice, Chicken...	17.0
122	Vomiting	[GERD, Chronic cholestasis, Peptic ulcer disease...	17.0
46	high_fever	[AIDS, Bronchial Asthma, Jaundice, Malaria, Ch...	12.0
72	Nausea	[Chronic cholestasis, Malaria, Dengue, Typhoid...	10.0
61	loss_of_appetite	[Chronic cholestasis, Peptic ulcer disease, Chi...	10.0
45	Headache	[Hypertension, Migraine, Paralysis (brain hem...	10.0
0	abdominal_pain	[Chronic cholestasis, Peptic ulcer disease, Jaundice...	9.0
131	yellowish_skin	[Chronic cholestasis, Jaundice, hepatitis A, H...	8.0
130	yellowing_of_eyes	[Chronic cholestasis, hepatitis A, Hepatitis B...	7.0
101	skin_rash	[Fungal infection, Drug Reaction, Chicken pox,	7.0

https://dx.doi.org/10.4314/aujst.v4i1.7

for an illness. We introduced feature selection to remove unwanted features using the importance threshold. The importance threshold is a value used to determine the level of importance for a feature. It is typically defined based on a certain quantile or percentile of the absolute values of the coefficients or feature importance scores. Finally, we are left with 3444 rows and 99 columns which we will use for our model training.

2.1.3 Generation of Synthetic Data using Synthpop:

Since the dataset is not enough to train our model, we needed to augment it. To expand the dataset for training, a surrogate dataset was generated based on the original data using the Synthpop library in Python. Synthpop generates synthetic data that closely resembles the characteristics of the original dataset while preserving the relationships between variables.

2.2 Experiment Setup

The study was carried out in Jupyter Notebook. To achieve the aim of this study, a set of experiments was conducted on the aforementioned dataset. The dataset was divided into a training set (80%) and a test set (20%) to obtain an optimal solution. The training set was utilized to train the proposed model, which subsequently made predictions for the test set. This division enabled the evaluation of the presented approach in terms of accuracy and performance. The proposed model was evaluated and compared with five other state of the art classifiers, K-Nearest Neighbours (KNN), Naive Bayes (NB), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and also Neural Networks (NN) have all been put through their paces in the context of experimental comparisons using Anaconda. Performance metrics include accuracy, Precision, Recall and F1-score.

2.3 Framework of the proposed model

2.3.1 Training the Proposed Model:

The proposed model was trained using both the original dataset and the surrogate dataset. The surrogate dataset was used to train the model while the original dataset was used for prediction. Starting from the left, the original dataset is inputted, then synthetic and encapsulated dataset is generated using the Synthpop library's generate () and Predictor_matrix() functions respectively. This generated dataset is then fed to the model for training. The model combined two algorithms, the weighted KNN algorithm and the rule-based system, to leverage their respective strengths and improve overall prediction accuracy.

2.3.2 Weighted KNN Algorithm:

The weighted KNN algorithm was employed in the proposed model. We use the Euclidean distance to calculate the distance, Value of k=2 to find nearest neighbours, Assign weights of 0.5. This algorithm is effective in making predictions for new data but can be sensitive to noise. To mitigate this, a distance-weighted voting mechanism was utilized. The neighbours' weights were determined based on the inverse of their distances, and the class labels were assigned based on the highest weighted vote. Weighted KNN can be sensitive

https://dx.doi.org/10.4314/aujst.v4i1.7

utilized to perform one-hot encoding on the

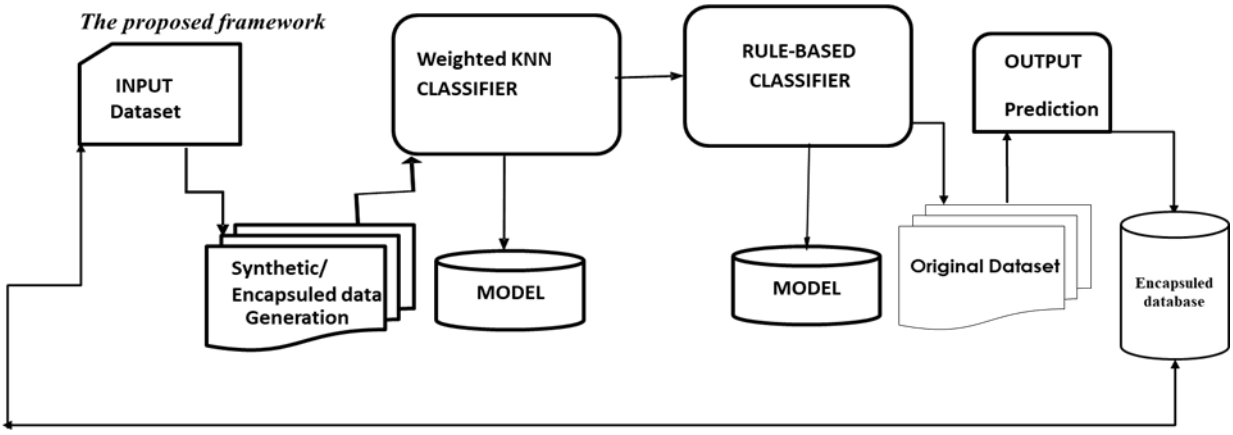


Figure 1: Proposed framework

Source: Das *et al.* (2020) adapted

to outliers or noisy data points in the dataset. Since the algorithm relies on distances between data points, outliers can have a significant impact on the weighting and influence the final predictions. To, Handle the noisy data, a rule-based algorithm was introduced.

2.3.3 Rule-Based System:

The rule-based system was integrated into the proposed model. This system is reliable in making predictions for data similar to the training data but can be brittle. If the rule-based system provided a specific output for a given instance, it was considered as the final prediction. Otherwise, the class label with the highest weighted vote from the weighted KNN algorithm was assigned as the final prediction.

2.3.4 Data Protection

As the model made predictions, the input data were stored directly into the database. The predictor function in the Synthpop library was data, ensuring the protection of patient-sensitive information. Additionally, this process made more data available to train the model further, enhancing its performance. The parameter used for the model are shown in Table 2.

Table 2: Parameters of the model and their values

Parameter	Value
k	2
Distance metric	Euclidean distance
Weight function	Inverse distance weighting
Normalize weights	False
Weights assigned	0.5

https://dx.doi.org/10.4314/aujst.v4i1.7

2.3.5 Algorithm

Algorithm: Weighted KNN with Rule-based algorithm

Begin

1. ***Step 1*** *-Let $L = \{ (x_i , y_i) , i = 1, \ldots ,n \}$ be a training set of observations x_i with given class y_i and let x be a new observation(query point), whose class label y has to be predicted.*

2. ***Step 2*** *- Compute $d(x_i, x)$ for $i = 1, \ldots ,n$, the distance between the query point and every other point in the training set.*

3. ***Step 3***- *Select $D' \subseteq D$, the set of k nearest training data points to the query points*

4. ***Step 4*** *-Predict the class of the query point, using distance-weighted voting. The v represents the class labels. Use the following formula*

5. ***Step 5***-*Calculate the weight for each neighbour based on the distance using the formula: weight = 1 / distance.*

6. ***Step 6*** *-Compute the total weight by summing up the weights of all neighbours.*

7. ***Step 7***-*Perform distance-weighted voting to predict the class label:*

a. *Initialize a dictionary class_votes_knn to store the class votes for each label.*

For each neighbour in D':
Calculate the weighted vote for the neighbour using the formula: weighted_vote = weight / total_weight.

Increment the vote count for the corresponding class label in class_votes_knn by adding the weighted_vote.

8. ***Step 8***- *If the output from the Rule-Based System is not None, assign the output to final_output.*

Otherwise, select the class label with the highest weighted vote from class_votes_knn and assign it to final_output.

End- *Return final_output as the predicted class label for the query point.*

Procedure:

Load the training data and test data.

Create a KNN model and a rule-base model.

Create a hybrid model that combines the KNN model and the rule base model.

Train the hybrid model on the training data.

Evaluate the hybrid model on the test data.

Repeat steps 4-5 for different weights assigned to KNN and rule-base algorithms.

https://dx.doi.org/10.4314/aujst.v4i1.7

2.4. Experiment Performance Metrics

We evaluated and compared the model with five machine learning classifiers namely; K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM) in terms of accuracy, precision, F1-score. Secondly, we train the model using an artificial neural network just to ensure the problem of overfitting is mitigated. The four quantitative evaluation measures used to evaluate the model's performance were accuracy, precision, recall and F1-score

Accuracy = (TN+TP)/(TN+TP+FN+FP)

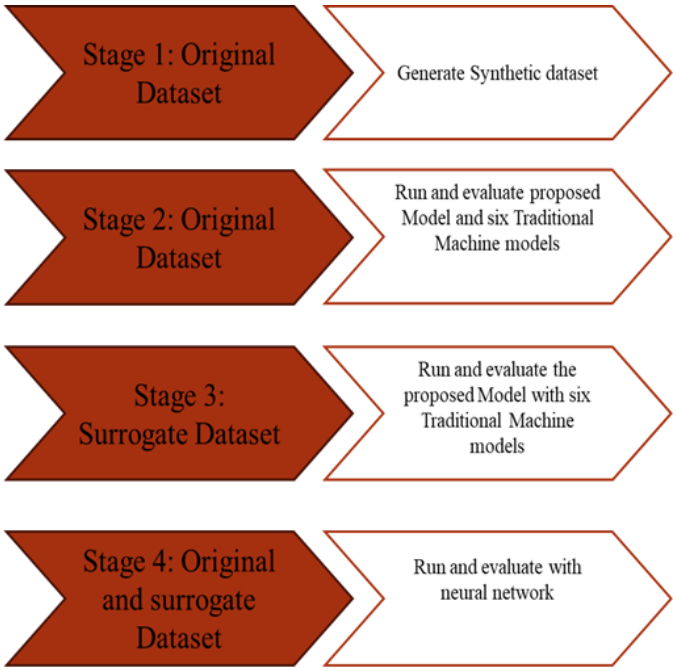
Precision = TP/ TP+FP

Recall = TP/ TP+FN

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

3. RESULTS AND DISCUSSIONS

3.1.1 Experiment Data Presentation:



3.1.2. Stage 1a:

In order to address the first objective of this paper, which focuses on proposing a mechanism for solving the dataset privacy problem in the Medical Diagnosis System, the following steps have been taken to ensure patient data privacy. To accomplish this, the predictor matrix function from the Synthpop library was used to transform all input variables into binary values of 0s and 1s. This transformation process, as demonstrated in Figure 5, effectively anonymizes the dataset while retaining its usability for training a medical diagnosis system. By converting the input variables into binary format, patient-specific information is protected, aligning with the goal of preserving data privacy in the context of the Medical Diagnosis System.

H_3: The proposed mechanism effectively protects patients' privacy in the Medical Diagnosis System.

3.1.2. Stage 1b

Secondly, to address the second objective of this paper of proposing a mechanism for solving data availability problems in the Medical Diagnosis System, the study uses the generator function in Synthpop library to generate synthetic data of 100,000 which imitates the original dataset of 4390 samples. Next, the training dataset and the testing dataset is combined which summed up to 5,300. Then this new dataset is used to generate the surrogate dataset of 100,000 samples. This dataset is now sufficient for a medical diagnosis system without any overfitting problem.

https://dx.doi.org/10.4314/aujst.v4i1.7

Table 3 : Records from the original database (Source: Research (2023)

	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
0	Umar Baba	back_pain	blood_in_sputum	cramps	enlarged_thyroid	extra_marital_contacts	AIDS
1	Abbas yahaya	abdominal_pain	chest_pain	distention_of_abdomen	fluid_overload	dischromic _patches	GERD
2	Tope Abdullah	blackheads	bruising	depression	distention_of_abdomen	extra_marital_contacts	GERD
3	Baba Umar	bruising	neck_pain	continuous_feel_of_urine	watering_from_eyes	muscle_weakness	Allergy
4	Yakubu Adamu	back_pain	depression	irritability	swelling_joints	loss_of_smell	Allergy
5	Nafiu Imam	abdominal_pain	family_history	history_of_alcohol_consumption	runny_nose	neck_pain	Allergy
6	Khamiz Faisal	altered_sensorium	dischromic _patches	diarrhoea	small_dents_in_nails	loss_of_smell	Fungal infection
7	Nazeefa Aliyu	excessive_hunger	dizziness	fluid_overload	rusty_sputum	muscle_pain	Hyperthyroidism

Table 4: Records from the new database. Source: Researcher (2023)

```
In [8]: spop.predictor_matrix
```

Out[8]:

	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
Name	0	0	0	0	0	0	0
Symtom1	1	0	0	0	0	0	0
Symtom2	1	1	0	0	0	0	0
Symtom3	1	1	1	0	0	0	0
Symtom4	1	1	1	1	0	0	0
Symtom5	1	1	1	1	1	0	0
Disease	1	1	1	1	1	1	0

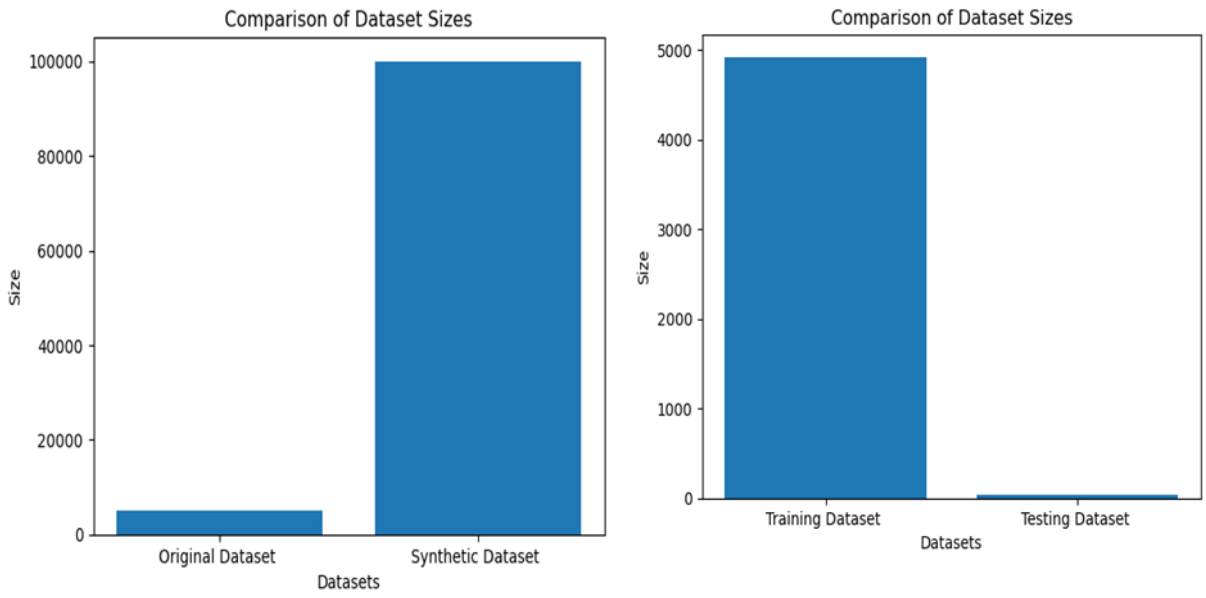


Figure 1: Distribution of the datasets, ource: Researcher (2023)

https://dx.doi.org/10.4314/aujst.v4i1.7

3.1.2. Stage 1c-

Here, the study tries to test if the surrogate dataset mimics the properties and attributes of the original dataset. To demonstrate that the surrogate dataset mimics the features of the original dataset, all properties of the synthetic data were compared to the original dataset on frequency plots (Figure 1). The histograms demonstrate the degree of similarity between the datasets is shown in Figure 2.

3.2. Model Training and Evaluations

The proposed model combines two models, the weighted KNN and Rule-based. This approach is a way to combine the strengths of both the Weighted KNN algorithm and the rule-based system. The model follows this step: Weighted Voting: When making predictions, the Weighted KNN algorithm employs a distance-weighted voting mechanism. The weights of the neighbours are determined based on the inverse of their distances. The class labels are assigned based on the highest weighted vote. Handling Rule-Based Output: If the rule-based system provides a specific output for a given instance, it is considered as the final prediction. Otherwise, the class label with the highest weighted vote from the Weighted KNN algorithm is assigned as the final prediction (Figure 6)..

3.2.1 Evaluation of the model with original dataset

We evaluate the proposed model with four other state-of-the-art traditional algorithms namely; Random Forest, Decision Tree, Naïve Bayes, K Nearest Nneighbor, Support Vector Machine and then use a Neural network to visualize the training process. Through this training, the training accuracies of the five classifiers were 100% while the testing accuracy for the five classifiers were between 0.2%-o.25%. This shows that our model is overfitting. Figure 7 shows the chart.

As we can see from the chart, The Proposed model classifier outperforms the other classifiers in terms of F1-score, suggesting that it achieves a better balance between precision and recall on the validation data compared to the rest of the classifiers. In summary, the best classifier is proposed model with scores of 98 for precision, 86 for recall, and 76 for F1-score, it signifies that the proposed model has the highest overall performance among the classifiers considered in terms of F1-score, while the other classifiers have comparable and slightly lower scores (Figure 8).

3.3 Evaluating with deep learning

3.3.1 Stage 2b-Train Artificial neural network with original dataset

To visualize the training process, an artificial neural network is trained. The model is trained for 10 epochs with SoftMax activation, Adam optimizer and categorical entropy. James et al. (2013, p13) suggest that overfitting can be identified through learning curves. When examining the learning curves, overfitting is indicated if the training loss continues to decrease as experience increases. Additionally, the validation loss initially decreases but then starts to increase again. The inflection point in the validation loss can serve as an indicator to halt the training process. In our case, the plot provided below clearly illustrates that our model is experiencing overfitting (Figure

https://dx.doi.org/10.4314/aujst.v4i1.7

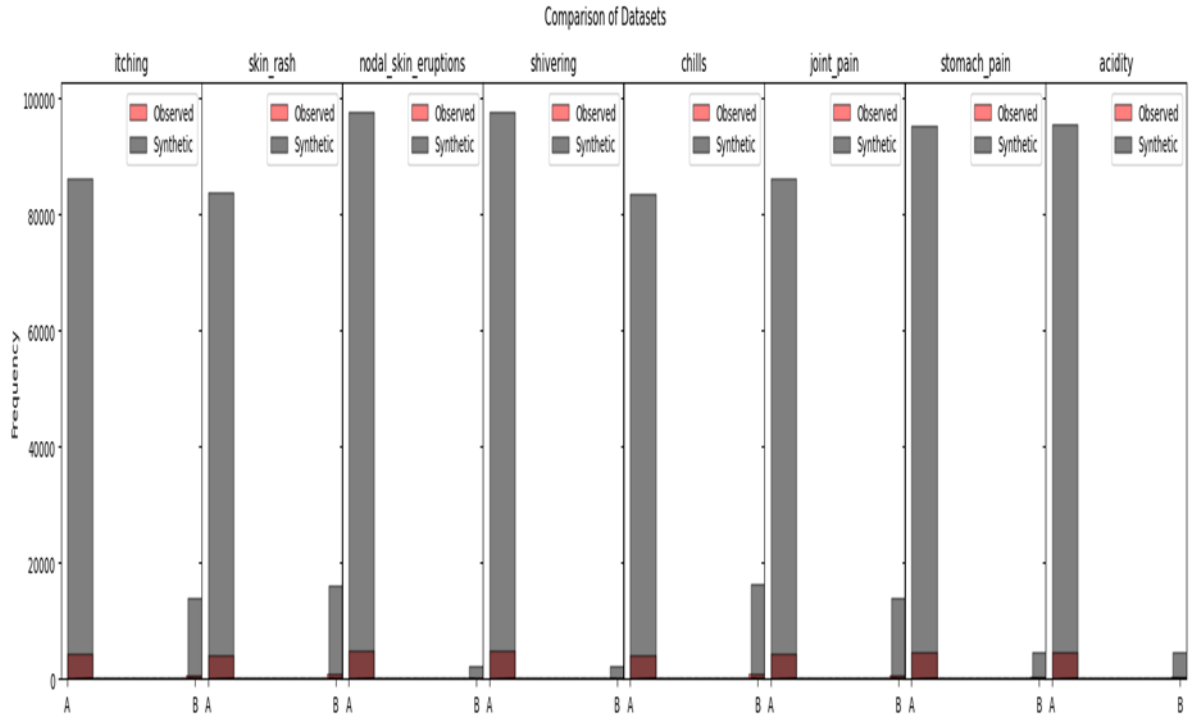


Figure 5: *Dataset Confidence Level*. Source: Researcher (2023)

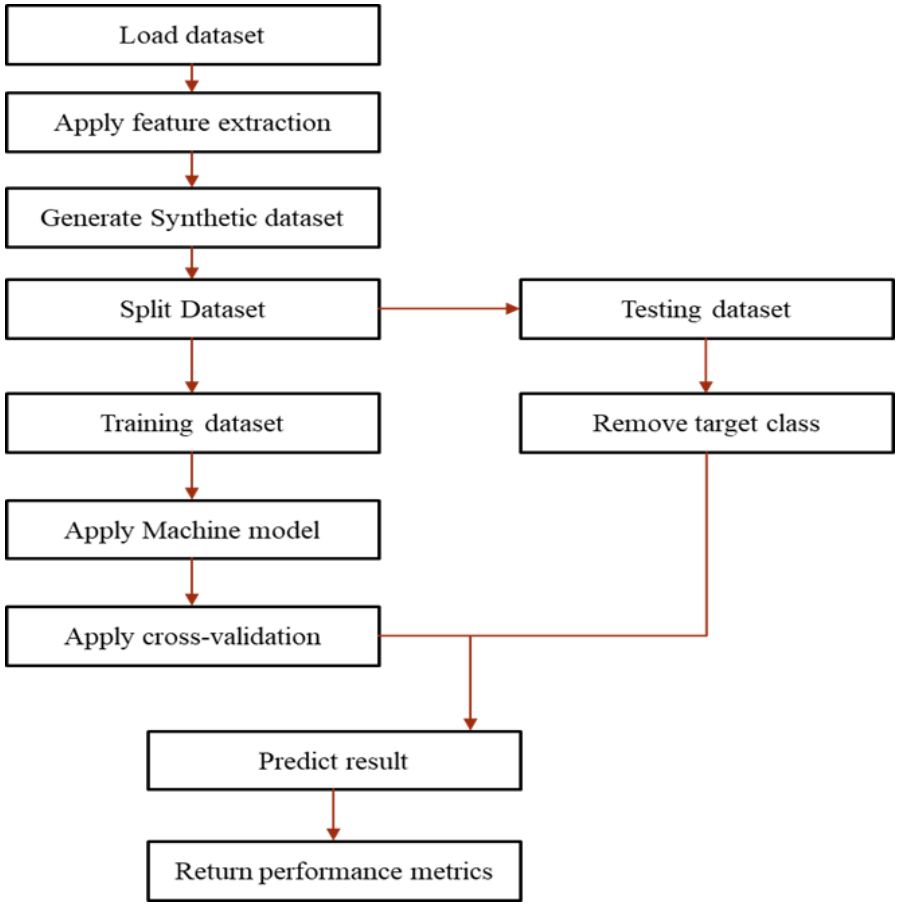


Figure 6: *Flowchart of the System*. Source: Researcher (2023)

https://dx.doi.org/10.4314/aujst.v4i1.7

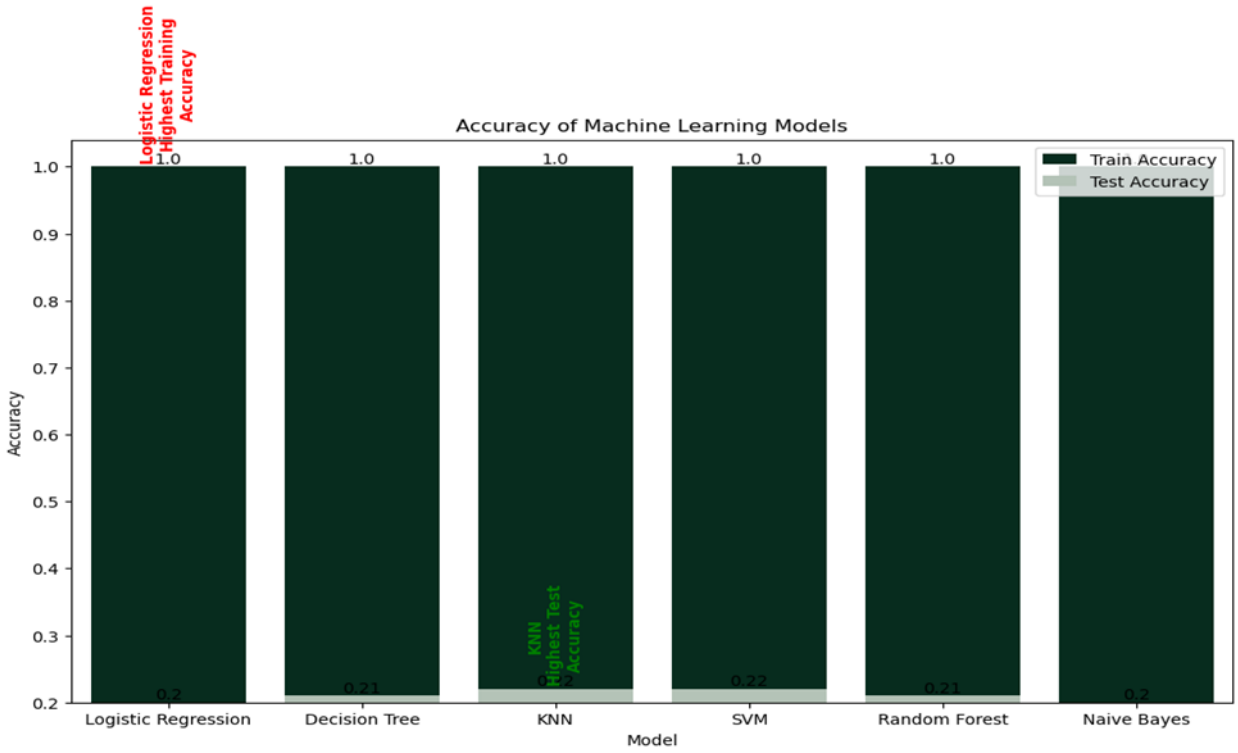


Figure 7: Training Accuracy vs Testing accuracy of five ML classifiers using original dataset

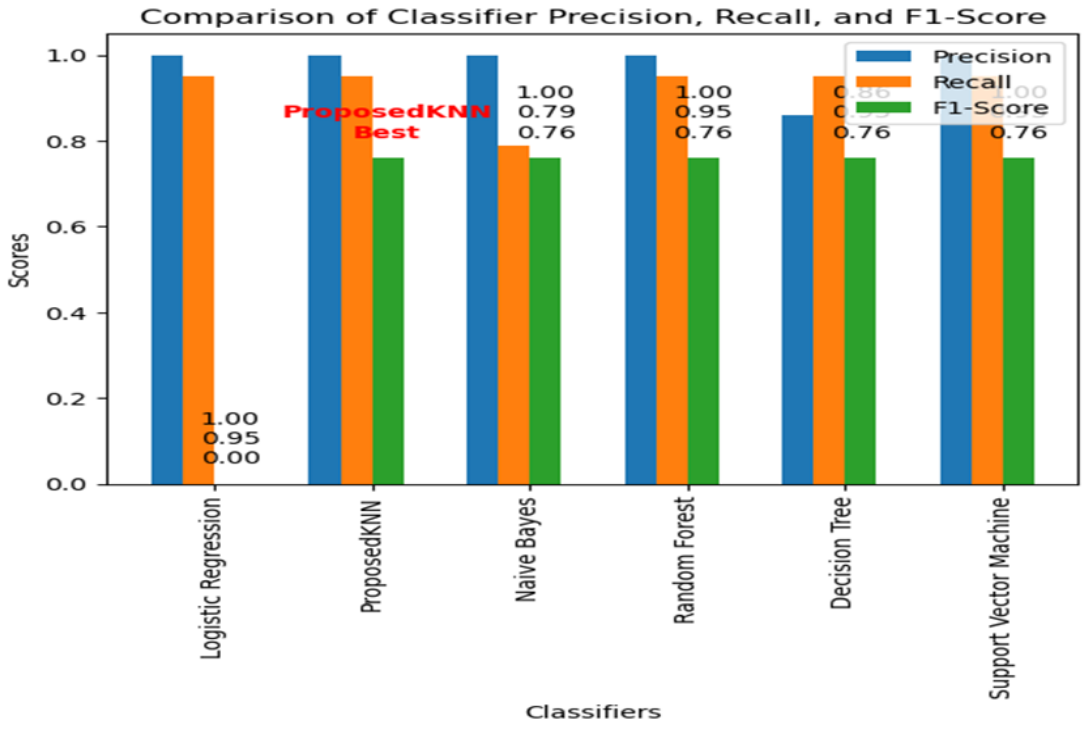


Figure 8: Comparison of Recall, Precision and F1-score of the five classifiers using original dataset

https://dx.doi.org/10.4314/aujst.v4i1.7

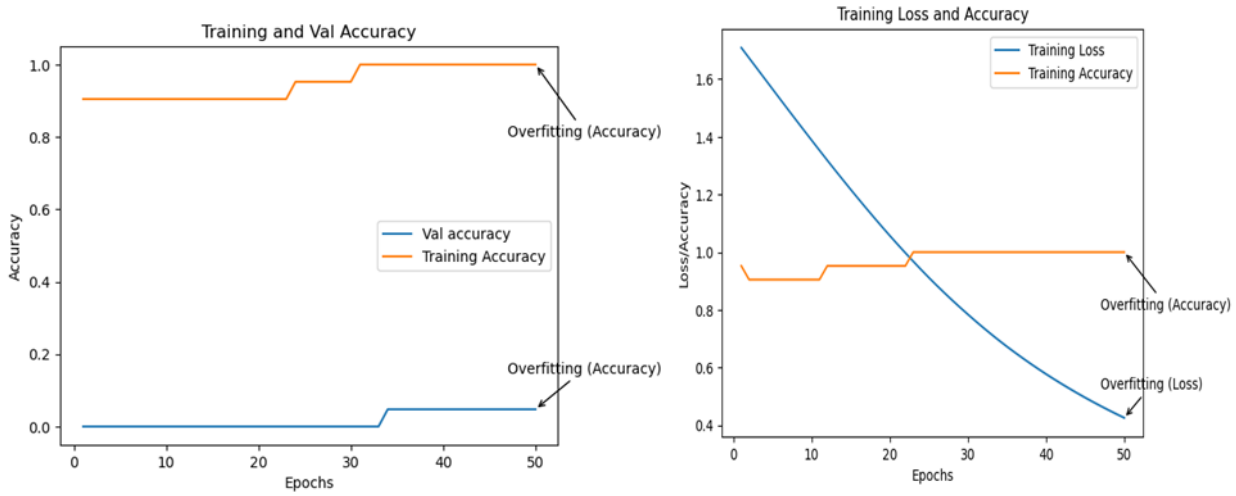


***Figure 9**: Learning graph using original dataset.* Source: Researcher (2023)

In this plot, the x-axis represents the number of epochs, and the y-axis represents the loss values. The blue line represents the training loss, while the red line represents the validation loss. If the training loss keeps decreasing, but the validation loss starts increasing or remains stagnant, it indicates overfitting. This study also looks at the training and Val accuracy and see that they are overfitting. This can commonly occur if the number of samples in a dataset is too small, relative to another dataset. There are two common cases that could be observed; they are: (i) Training dataset is relatively unrepresentative. (ii) Validation dataset is relatively unrepresentative.

3.4 Evaluating the model with Synthetic dataset

3.4.1 Stage 3a- Training the model with the Synthetic dataset and comparing them with five classifiers and the proposed Model. The model is trained with the Synthetic dataset using Random Forest, Decision Tree, Naïve Bayes, SVM, and the Proposed Model, the training accuracies of the five classifiers were on the range 0.994%-0.998% while the testing accuracy for the four classifiers have now increased to 0.996%-0.997%. This shows that our model is no longer overfitting and can now generalize. Figure 10 shows the chart.

All classifiers have precision, recall, and F1-score of 98%, which indicates that they are performing very well in terms of accurately predicting positive samples (precision), capturing true positive samples (recall), and achieving a balanced measure of precision and recall (F1-score).

3.5 Evaluating the model using deep learning with Surrogate dataset

3.5.1 Stage 3b- Training using Artificial Neural Network- The model is now trained using the synthetic dataset of 100,000 samples. The model is trained for 50 epochs. According to James et al., (2013) A good fit can be observed in learning curves when the training loss decreases and stabilizes, and the validation loss also decreases and reaches a stable point with a small difference compared to the training

https://dx.doi.org/10.4314/aujst.v4i1.7

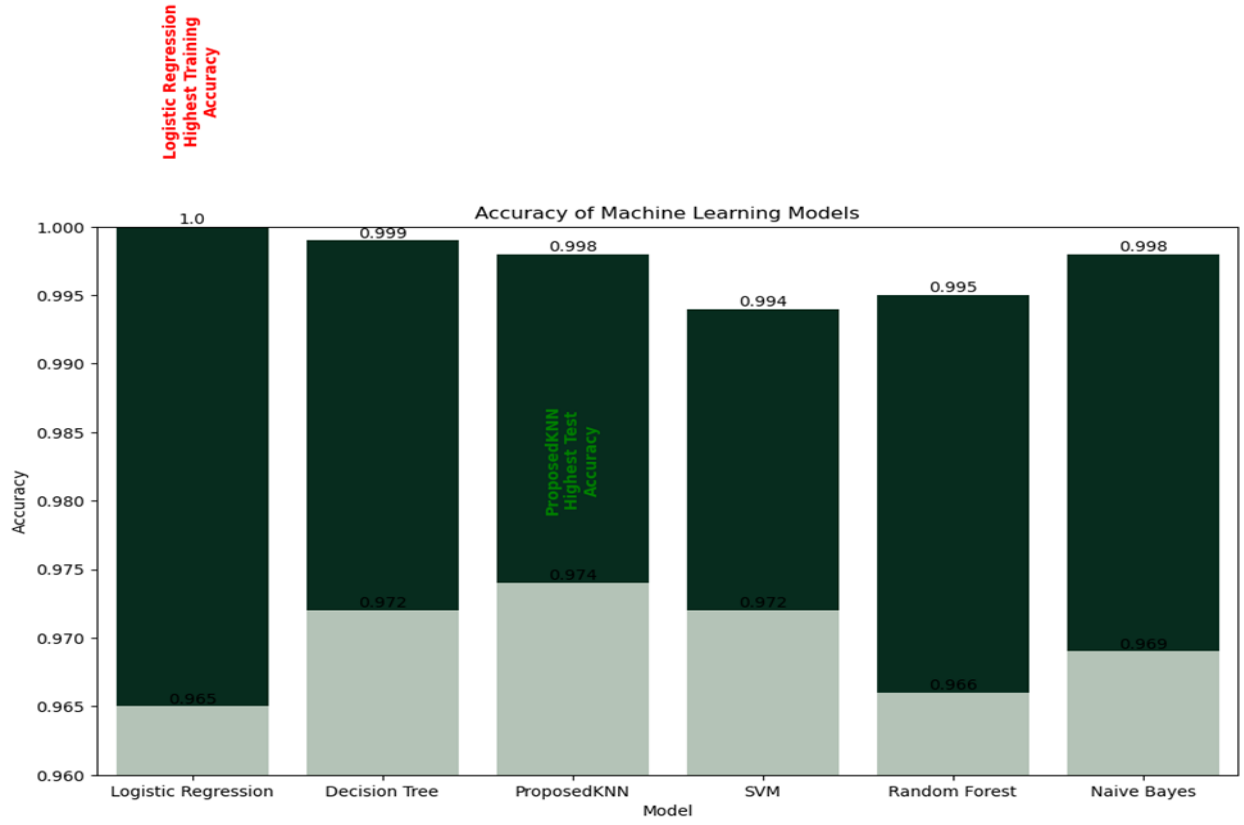


Figure 10: *Training and Testing accuracy graph of synthetic dataset.*

H_5: The proposed system achieves better accuracy than start-of the-classifiers

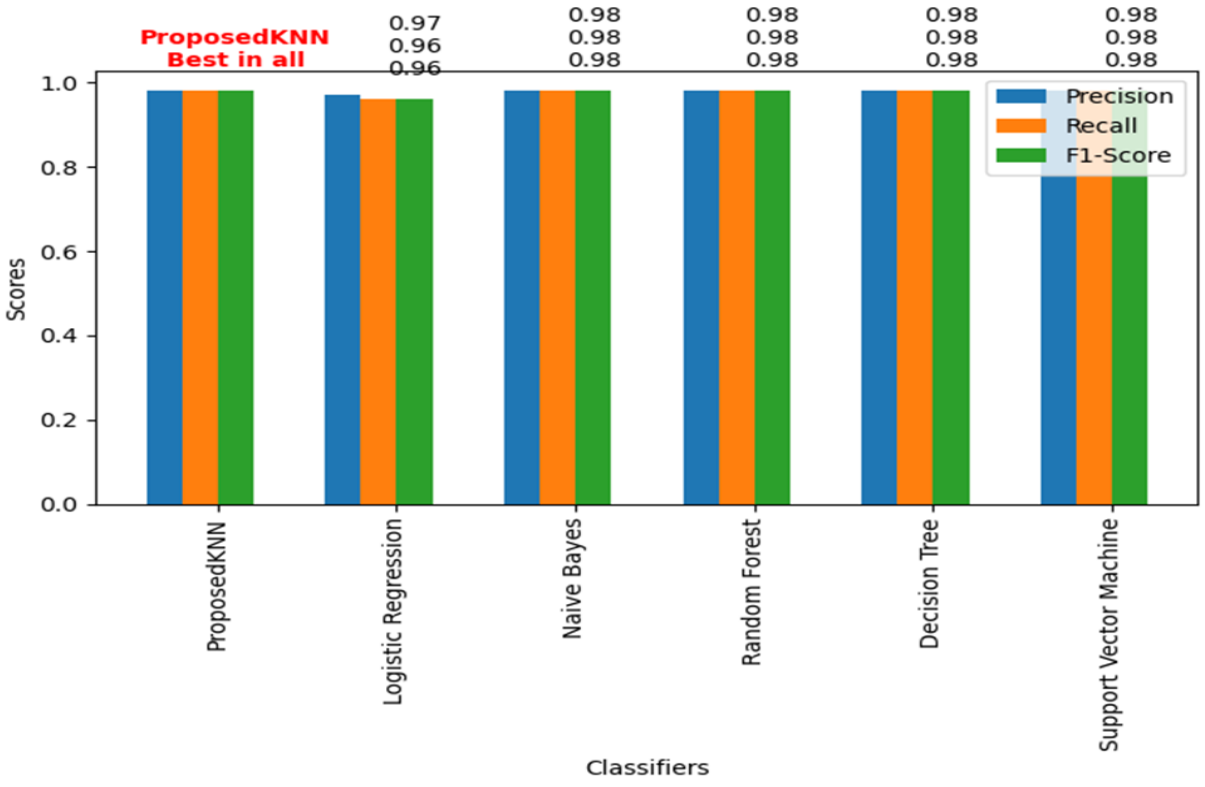


Figure 11: *graph showing the Precision, recall and F1-score for each classifier*

https://dx.doi.org/10.4314/aujst.v4i1.7

accuracies of the five classifiers were on the range 0.994%-0.998% while the testing accuracy for the four classifiers have now increased to 0.996%-0.997%. This shows that our model is no longer overfitting and can now generalize. Figure 10 shows the chart.

All classifiers have precision, recall, and F1-score of 98%, which indicates that they are performing very well in terms of accurately predicting positive samples (precision), capturing true positive samples (recall), and achieving a balanced measure of precision and recall (F1-score) (Figure 11).

3.6 Evaluating the model using deep learning with Surrogate dataset

3.6.1 Stage 3b- Training using Artificial Neural Network- The model is now trained using the synthetic dataset of 100,000 samples. The model is trained for 50 epochs. According to James et al., (2013) A good fit can be observed in learning curves when the training loss decreases and stabilizes, and the validation loss also decreases and reaches a stable point with a small difference compared to the training loss. If training is continued beyond this point, it is likely to result in overfitting. The plot below demonstrates our model is now a good fit.

In this plot, the x-axis represents the number of epochs, and the y-axis represents the loss values. The blue line represents the training loss, while the red line represents the validation loss. The training loss keeps decreasing, and the validation loss remains stagnant, it indicates not overfitting.

H_4: The proposed mechanism effectively solves the overfitting problem in training the Medical Diagnosis System.

4. Summary of Experiment Result

The results of this study have been validated through rigorous analysis and comparison. Various validation techniques were employed to ensure the accuracy and reliability of the findings. These validation methods include in Tables 3 and 4.

5. CONCLUSION

In conclusion, this paper has presented a holistic solution to address the challenges of data privacy and data availability in the field of medical diagnosis. By combining innovative techniques in data generation, data encapsulation, and a hybrid approach of the weighted KNN algorithm and a rule-based algorithm, we have proposed a comprehensive framework that aims to safeguard patient privacy while ensuring accurate and reliable diagnostic predictions.

Through the use of synthetic data generation, we have demonstrated the ability to preserve the statistical properties of the original dataset, thereby protecting sensitive patient information. This approach not only addresses the privacy concerns associated with sharing medical data but also provides a valuable resource for training and testing medical diagnosis systems without compromising patient confidentiality.

Furthermore, the integration of the weighted KNN algorithm and the rule-based algorithm

https://dx.doi.org/10.4314/aujst.v4i1.7

Table 3: Summary of Result with Original dataset ((Source: Researcher,2023)

	LR	RF	DT	NB	SVM	ANN	Proposed Model
Train Accuracy	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Test Accuracy	0.21	0.22	0.21	0.22	0.22	0.21	0.23
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.95	0.95	0.95	0.76	0.95	0.96	0.79
F1-score	0.79	0.76	0.76	0.76	0.76	0.76	0.79

Table 4: Summary of training with surrogate dataset

	LR	RF	DT	NB	SVM	ANN	Proposed Model
Train Accuracy	1.00	1.00	1.00	1.00	1.00	0.98	1.00
Test Accuracy	0.98	0.98	0.98	0.98	0.98	0.98	0.99
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.95	0.95	0.95	0.76	0.95	0.96	0.79
F1-score	0.79	0.76	0.76	0.76	0.76	0.76	0.79

Table.5: Summary of original and surrogate dataset

	Train-Acc	Test-Acc	Precision	Recall	F1-score
Proposed Model with original Dataset	1.00	0.23	1.00	0.79	0.79
Proposed Model with surrogate Dataset	1.00	0.98	0.98	0.98	0.98

https://dx.doi.org/10.4314/aujst.v4i1.7

offers a robust prediction mechanism. While the weighted KNN algorithm excels in making predictions for new data, its sensitivity to noise is mitigated by the rule-based algorithm, which incorporates expert knowledge and well-defined rules to improve the overall prediction accuracy. This hybrid combination leverages the strengths of both approaches, resulting in a more robust and reliable diagnostic system. The experimental evaluations conducted in this study have showcased the effectiveness and superiority of our proposed methodology compared to existing approaches. The comprehensive performance metrics, including accuracy, precision, recall, and F1-score, highlight the enhanced predictive capabilities and privacy preservation achieved through our framework.

REFERENCES

Aswal, S., Ahuja, N., & Ritika. (2016). Experimental analysis of traditional classification algorithms on bio medical dataset. *Experimental analysis of traditional classification algorithmInternational Conference on Next Generation Computing Technologies (NGCT)*, 566-568.

Ateniese, G., Fu, K., Green, M., & Hohenberger, S. (2020). Privacy-Preserving Medical Data Sharing and Analysis *Applied Sciences*, *12*(23), 12320. https://doi.org/10.3390/app122312320

Bergstra, James & Bengio, Y.. (2012). Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research. 13. 281-305.

Chen, Y., Torkzadehmahani, R., Nasirigerdeh, R., Blumenthal, D. B., Kacprowski, T., List, M., Matschinske, J., Spaeth, J., Wenke, N. K., & Baumbach, J. (2022). Privacy-Preserving Artificial Intelligence Techniques in Biomedicine. Methods of information in medicine, 61 (S 01), e12–e27. https://doi.org/10.1055/s-0041-1740630.

Collins, J., Brown, J., Christine, S., Hutson, K., & Jeffery, E. (2017). Meaningful Analysis of Small Data Sets: A Clinician's Guide. *Greenville Health System Proc, 1*, 16-19.

Das, H., Naik, B., & Bahera, H. (2020). Medical disease analysis using Neuro-fuzzy with feature Extraction Model for classification. *Informatics in Medicine Unlocked*, 1-12. doi: 10.1016/j.imu.2019.100288

El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, E. M. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. *Procedia Computer Science, 65,* 459-468.

HIPAA. (2013). HIPAA Privacy Rule. *45 CFR*. doi:160.103 2013.

Hoang, L., Manh, T., Fujita, H., Dey, N., Ashour, A. S., Truong, V., Ngoc, N., Quynh, L., & Chu, D. (2018). *Biomedical Signal Processing and Control Dental diagnosis from X-Ray images: An expert system based on fuzzy computing*. *39*, 64–73.

Imhanlahimi, R., & Otumu, J. (2019). Application of Expert System for Diagnostic Medical Conditions; A Methodological Review. *European Journal of Computer Science and Information Technology, 7*(2),

https://dx.doi.org/10.4314/aujst.v4i1.7

I7(2), 12-25.

Kononenko, B. I., & Kukar, M. (2020). Application of machine learning to medical diagnosis. *machine learning and data mining: methods and Applications, 389*, 408

Marsland, S. (2015). Machine learning: an algorithmic perspective. *CRC press*.

Nnebe, S. E., Okoh, N. A., John-Otumu, A., & Osaze, E. (2019). A Neuro-Fuzzy Case Based Reasoning Framework for Detecting Lassa fever Basedon Observed Symptoms. (SciencePC, Ed.) *American Journal of Artificial Inteligence, 3*(1), 9-16.

Nowok, B., Raab , G., & Dibben, C. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, 74*, 1-26. Retrieved from Synthpop: Bespoke Creation of Synthetic Data in R", Journal ohttps://www.jstatsoft.org/v074/i11

Patra, S., Sundar, G., & Thakur, M. (2014). *A Proposed Neuro-Fuzzy Model for Adult Asthma Disease Diagnosis A P ROPOSED N EURO -F UZZY M ODEL F OR. March 2013*. https://doi.org/10.5121/csit.2013.3218

Rana, M., & Srdamkar, R. R. (2018, June). Design of expert system for medical diagnosis using fuzzy logic. *International Journal of Scientific and Engineering Research, 4*(6), 2914-2921.

Sabay, A., Bejugama, V., & Jaceldo-siegl, K. (2020). Overcoming Small Data Limitations in Heart Disease Prediction by using Surrogate Data.

Torgyn, S., & Khovanova, N. (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine, 75*, 51-63.

Tao, M., Yang, Y., & Li, D Ji-Jiang Yang, Jian-Qiang Li, and Yu Niu. 2015. A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Gener. Comput. Syst. 43, C (February 2015), 74–86. https://doi.org/10.1016/j.future.2014.06.004

Qiu, W., Dong Li, Xiaofeng Liao, Tao Xiang, Jiahui Wu, and Junqing Le. 2020. Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. Computer Security 90, C (Mar 2020). https://doi.org/10.1016/j.cose.2019.101701.