**A Comparative Study of Machine Learning for Predicting Multiple Diseases**

This research is presented to the Department of Computer Science and Engineering at Jahangirnagar University as a partial fulfillment of the requirement for the degree of Bachelor of Science and Engineering.

***Submitted By***

**Mehadi Hasan**

Exam Roll: 180696

Registration No: 46341

Session: 2017-2018

**Sohanur Rahman**

Exam Roll: 180703

Registration No: 47939

Session: 2017-2018

***Supervised By***

**Dr. Israt Jahan**

Professor

Department of Computer Science and Engineering

Jahangirnagar University.



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAHANGIRNAGAR UNIVERSITY.

SAVAR, DHAKA 1342.

MAY-2023

**Abstract**

Our everyday lives' most crucial aspect is how well we are mentally and physically. Heart disease, diabetes, and pneumonia are merely a few of the health issues that have recently become more common in our daily lives and in the field of healthcare. And this expenses humanity an enormous quantity of time. Based on the available data, machine learning can be used to detect such situations. The work, which is called "Multiple Disease Prediction" attempts to demonstrate how machine learning can be used to model the collection of data. The model is then applied to identify if a person is suffering from the disease or not. We employed a variety of methods, including decision trees, K-nearest neighbors (KNN), support vector machines (SVM), naïve bayes, and linear regression. Results of these algorithms are compared using their accuracy, precision, recall, and F1-score. The confusion matrix is used to plot the ROC curve. The technique with the best accuracy, precision, recall, and F1-score is taken into consideration for determining the optimum algorithm for illness detection after these algorithms are compared for accuracy, precision, recall, and F1-score.

# Declaration

The project work **"A Comparative Study of Machine Learning for Predicting Multiple Disease"** iscompleted at Department of Computer Science and Engineering, Jahangirnagar University is unique and conforms to the university's regulations.

We are aware of the University's plagiarism policy and certify that no component of this project has been plagiarized or previously submitted for the granting of any degree or diploma.

------------------------------------- -----------------------------------------

**Mehadi Hasan** **Sohanur Rahman**

Exam Roll: 180696 Exam Roll: 180703

Registration No:46341 Registration No: 47939

Session: 2017-18 Session: 2017-18

--------------------------------------------

**Dr. Israt Jahan**

Professor

Department of Computer Science and Engineering

Jahangirnagar University

# Acknowledgement

All thanks and praise are due to Allah, who made it possible for us to complete this project successfully with the help of his heavenly blessing.

We are grateful to Dr. Israt Jahan, our esteemed supervisor, who is a professor in the department of computer science and engineering at Jahangirnagar University, for her patient counsel, astute direction, insightful instruction, and creative suggestion throughout the project's duration. This undertaking would not be possible without her invaluable aid. She has consistently been a source of inspiration and encouragement for us to put in a lot of effort.

Finally, we had wanted to give a shout-out to all of our friends who are truly dear to our hearts. We will never be willing to find the perfect words to express our gratitude to our loving parents, who have committed moral support and fortification in the completion of the project.

**Contents**

1. **Introduction 1**
2. **Related Work 2**
3. **Problem Definition and Algorithm 3**

2.1 Task Definition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.2 Algorithm Definition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2.1 Decision Tree Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2.2 Support Vector Machine Algorithm . . . . . . . . . . . . . . . . . . . . . . . 5

2.2.3 Logistic Regression Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

* + 1. Convolutional Neural Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

1. **Experimental Evaluation 8**
   1. Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

3.1.1 Experimental Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

3.1.2 Performance Evaluation Metrics . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19  
  
 3.3 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

**5 Future Work 28**

**6 Conclusion 29**

**List of Figures**

**1.1** Diagram of a Credit Card Fraud Detection Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

**3.1** Approached Credit Card Fraud Detection System . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

**3.1.1.1** Sample Dataset for Credit Card Fraud Detection Model . . . . . . . . . . . . . . . . . . . . . . . 10

**3.1.1.2** Graphing the Age of Credit Card by Gender . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**3.1.1.3** Frequency Distribution of Credit Card’s Customer Gender . . . . . . . . . . . . . . . . . . . . . 12

**3.1.1.4** Gender Distribution of Customers as a Pie Chart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

**3.1.1.5** Frequency Distribution of Credit Card Age . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

**3.1.1.6** Age Distribution of Credit Card as a Pie Chart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

**3.1.1.7** Visual Illustration of Two Outputs with the Numbers 0 and 1

Indicating Whether They Are Legitimate or Fraudulent . . . . . . . . . . . . . . . . . . . . . . . 15

**3.1.2** A 22 Confusion Matrix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

**3.2.1** Classification Report for Decision Tree Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

**3.2.2** ROC Curve for Decision Tree Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

**3.2.3** Classification Report for K-Nearest Neighbor Algorithm . . . . . . . . . . . . . . . . . . . . . . . . 20

**3.2.4** ROC Curve for K-Nearest Neighbor Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

**3.2.5** Classification Report for Support Vector Machine Algorithm . . . . . . . . . . . . . . . . . . . 21

**3.2.6** ROC Curve for Support Vector Machine Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . 22

**3.2.7** Classification Report for Naive Bayes Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

**3.2.8** ROC Curve for Naive Bayes Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

**3.2.9** Classification Report for Logistic Regression Algorithm . . . . . . . . . . . . . . . . . . . . . . . 23

**3.2.10** ROC Curve for Logistic Regression Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

**List of Table**

**3.3** A comparative result of our proposed model using machine learning algorithm **25**

**Chapter 1**

**Introduction**

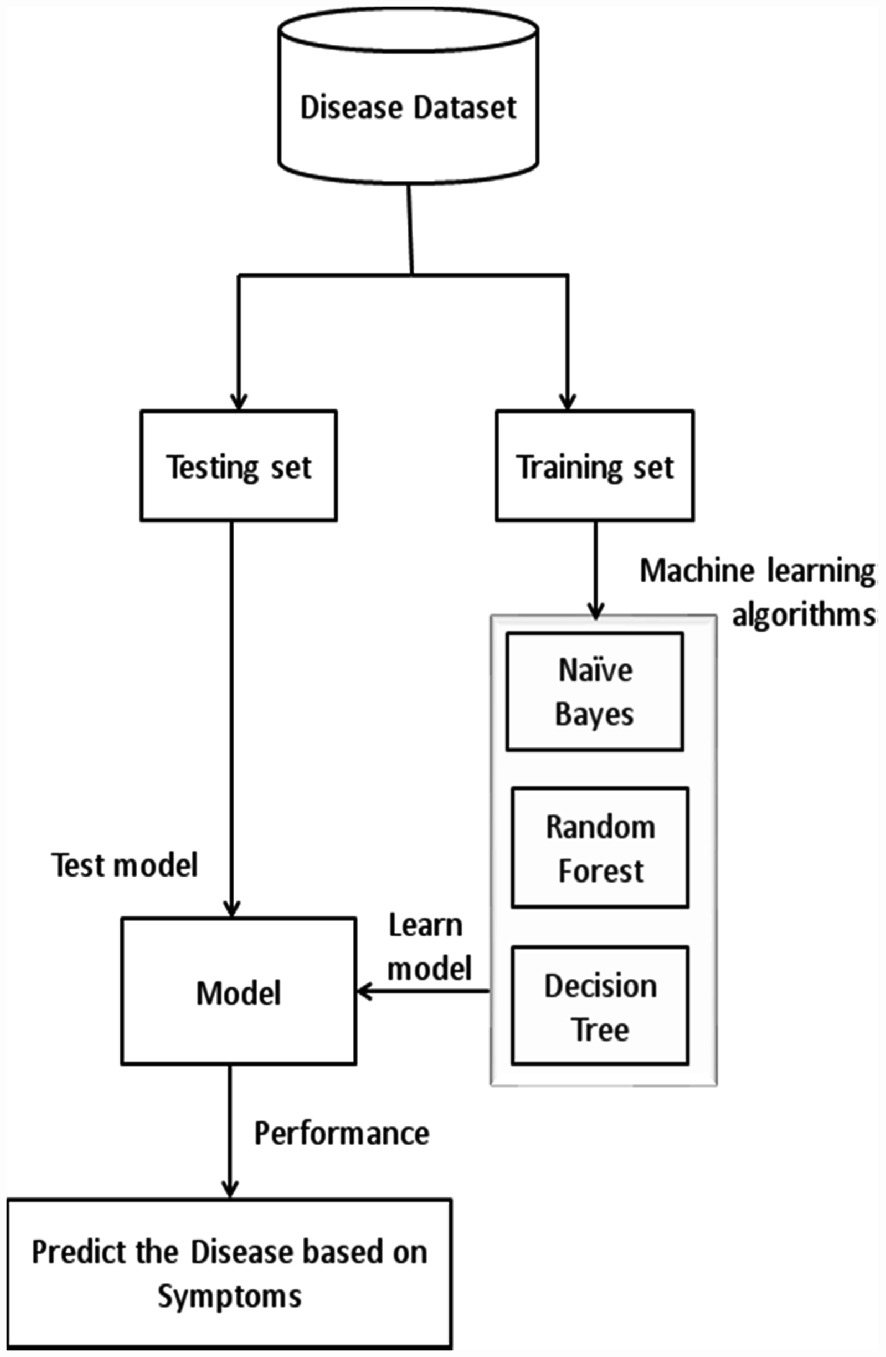
The goal of the rapidly expanding field of healthcare known as disease detection using machine learning is to create accurate and effective algorithms for diagnosing diseases based on patient data. It is now possible to leverage a lot of data to create models based on machine learning for disease identification, thanks to the expanding availability of electronic health records, medical imaging, and genetic data. On the basis of this data, machine learning algorithms may be trained to identify patterns and connections between the characteristics of the diseases and the attributes. These algorithms can then be used to forecast whether a disease will manifest in a new patient.

Machine learning-based identification of illnesses has considerable potential advantages. Machine learning models can assist healthcare professionals in making more precise and swifter diagnoses, which is essential for the effective treatment of diseases. Additionally, risk indicators for diseases can be found, and particular to the patient treatment plans may be developed using machine learning.

Data quality, data privacy, model interpretability, and model generalization are just a few of the difficulties that come with utilizing machine learning to detect diseases. The ability to overcome these obstacles and create efficient disease detection systems is now possible because to developments in machine learning algorithms and methods. In light of this, disease identification via machine learning is an interesting area of research that has the potential to revolutionize healthcare and enhance patient outcomes.

This study can teach us about a reliable feedback machine learning based medical disease detection system. With the aid of this feedback technique, the classifier's detection rate and performance are enhanced. Examine the performance of a number of classification algorithms, such as Decision Tree (DT) classifiers, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), and Linear Regression methods, on a completely unbalanced medical disease detection database.

In our research to identify numerous diseases, the proposed model is visually shown. The dataset for a person's disease is read by our algorithm using data downloaded from Kaggle.

****

**Figure 1.1:** Diagram of multiple disease detection model

This is the approach we suggest using to assess the model.

**Chapter 2**

**Problem Definition and Algorithm**

**2.1 Task Definition**

It may end up in many different kinds of health problems in the future if a particular individual has been experiencing a few symptoms but is unaware of the illness they are experiencing. This disease prediction will be highly helpful to a variety of people, including children, teens, adults, and elderly citizens, in order to prevent this and learn about the condition in the very beginning stages of the symptoms. In order to acquire the suitable treatment needed to address the condition, the person might adopt preventive measures or seek expert medical guidance.

The scope of a disease prediction system is particularly broad, illustrating how the world continues to evolve and how advancements in technology bring with them numerous disadvantages, which include a variety of food adulterations, inadequate nutrient supply to the body, unhealthy lifestyles involving improper consumption of food, as well as issues like obesity or unhealthy weight. Many different diseases go along with all of this.

Offering the finest quality services to all patients in the medical or healthcare fields is a significant task, and only those who can afford it can profit from it. There is a sizable amount of healthcare data that is not being mined in a more trustworthy and effective way to find hidden information for making good decisions. Methods based on data mining are used in the proposed framework to find chronic diseases early. Programming computers to come up with better results based on examples or past data has become known as machine learning. Machine learning pertains to the study of computer systems that learn from information and experience. The predictive machine learning algorithm includes two stages: training and testing.

Unfortunately, people often overlook their health because they are too preoccupied with their everyday tasks. Children and elderly people are both capable of ignoring or failing to recognize the crucial indicators that can later lead to more serious problems. It is advisable to get treatment before the illness worsens and progresses. Such individuals can benefit from preventative care and early detection of their health conditions with the use of a prediction system. This facilitates access to primary healthcare in isolated regions.

It is feasible to anticipate more than one disease at once when using the multiple disease prediction method. In order to anticipate the ailments, the user does not have to visit many sites. We are focusing on the disorders of the lung, diabetes, and the heart. because the three illnesses are related to one another. We're going to use Streamlit and machine learning methods to implement multiple illness analyses. When a user accesses this API, they must send the disease's parameters as well as the name of the disease. The appropriate model will be called by Streamlit, which then delivers the patient's state.

**2.2 Algorithm Definition**

An algorithmic and mathematical framework known as machine learning enables a computer system to learn from its prior experiences without being explicitly instructed what to do. Learning is concerned with the mimicking of human behavior by computers as well as the improvement of learning via the use of historical data. It also wants to create a data-driven system that is more predictive and adaptable. Here are some illustrations of machine learning techniques.

**2.2.1 Decision Tree Algorithm**

An example of supervised learning [1] is a decision tree, which can be applied to problems like classification and regression. It has the capacity to work with numerical and discrete data. It has a tree-like design with nodes and branches beginning at the tree's base and extending on subsequent branches till reaching the leaf node. The central node symbolizes the properties of the dataset, the branches the properties of the rules, and the leaf nodes the qualities of the solution to the issue. Decision tree algorithms are used in the real world for things like differentiating between cancerous and non-cancerous cells and giving customers car-buying recommendations.

There are several different data mining methodologies, which include ID3, CART, J48, NB Tree, REP Tree, and others. A broad algorithmic approach that has been often utilized to construct classification models is a tree structure. Most decision tree induction approaches employ a greedy top-down recursive partitioning methodology for tree development.

**2.2.2 Support Vector Machine Algorithm**

A supervised learning method for regression and classification issues is the support vector machine (SVM). However, it is primarily employed to address categorization issues. SVM is used to establish a decision boundary or collection of points, that categorizes data.

Because support vectors are the data sets that help define the higher dimensional space, the technology is known as a support vector machine. SVM may be utilized for a variety of tasks, including medication development, picture classification, and face identification.

**Pseudo code:**

• Importing all of the necessary packages

For example, import pandas as pd.

• SVM defense

**STEP-1:** Start

**STEP-2:** Reading the dataset # reads the pd.read.csv dataset (file name)

**STEP-3:** Cleaning and preparation of data. Relevant Data is resized as normal and fraudster classes, with normal = 0 and fraudster = 1 in the normal and fraudster classes, respectively.

• Data is under saturated;

• Data is scaled (null values are deleted); and data is normalized

• Using the split () function on the training phase, the data is split into two sets: training dataset and testing dataset.

**Step-4:** Train the dataset with the Support Vector Machine algorithm.

• classifier.predict () # is an SVM classifier that forecasts whether or not a transaction is fraud.

**Step-5:** Computing the number of fraudulent and genuine transactions, and also recall, precision, and accuracy, and placing the findings in the proper places.

**STEP-6:** Stop

**2.2.3 Linear Regression Algorithm**

The link between a dependent variable and one or more independent variables may be modeled statistically using linear regression. It presumes that the variables have a linear connection, which means that changes to one or more of the independent variables will result in proportionate changes to the dependent variable.

In basic linear regression, there is just one independent variable, and a straight line is used to represent the connection. Finding the best-fitting line that illustrates the connection between the variables is the aim of linear regression. To do this, the difference between the observed values of the dependent variable and the expected values based on the independent variable is squared, and the difference is minimized.

When there are several independent variables, multiple linear regression is an extension of simple linear regression. In this instance, a hyperplane in multidimensional space is used to model the connection.

In several disciplines, including economics, finance, engineering, and social sciences, linear regression is often employed. In addition to determining the degree and direction of the link between variables, it is frequently used for forecasting and prediction. Numerous machine learning algorithms, including neural networks and support vector machines, are based on linear regression.

**2.2.4 Convolutional Neural Network**

A sort of artificial neural network called a convolutional neural network (CNN) is made to do tasks like image and video recognition. They take their cues from the organization and operation of the brain's visual cortex, which employs a hierarchical approach to processing visual data.

Instead of requiring manual feature engineering, CNNs are made to automatically learn and extract features from images. Convolutional layers, which apply a collection of learnable filters (sometimes referred to as kernels) to the input image to produce a set of feature maps, are used to do this. These feature maps identify the presence of particular textures, edges, and other patterns and structures in the input image.

Pooling layers, which down-sample the feature maps by taking the maximum or average value in a particular region, are also frequently included in CNNs. By doing so, the feature maps' size can be shrunk and the most important features can be extracted.

**Chapter 3**

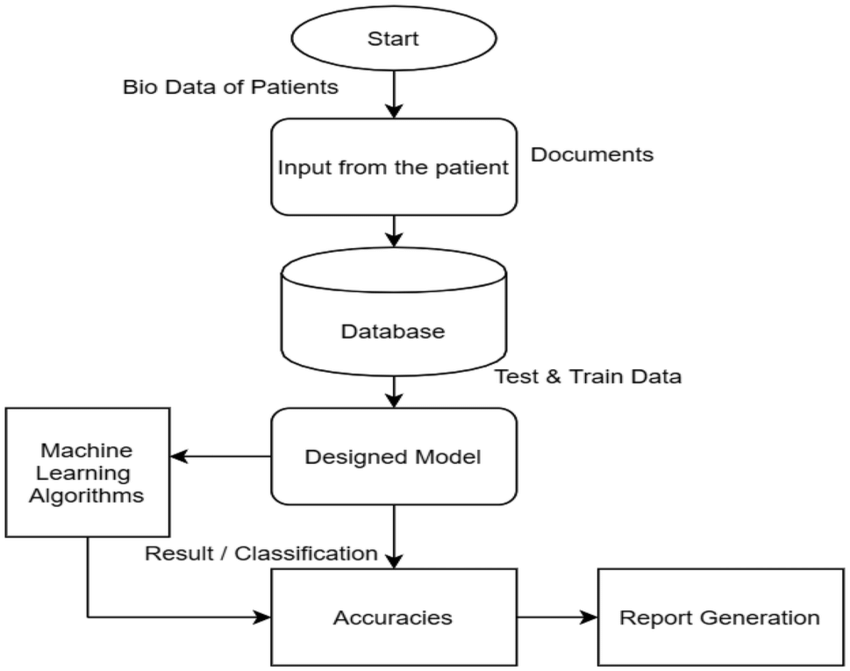
**Experimental Evaluation**

This section discusses the research process, findings, and conclusions from our investigation of numerous illness prediction systems. Because of its ability to handle both numerical and image data, we chose this assessment method when gathering our data. Three distinct forms of data will be included because of the three different types of research we have done. During the application step, the data set was divided into two groups. These two categories are, respectively, the training set and the testing set. We used a total of 25% of the data for testing and 75% of the data for training in order to produce a more effective model. The model was built using machine learning techniques such as Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR) and Convolutional Neural Network (CNN). The best at identifying the abnormality was the Logistic Regression.

**3.1 Methodology**

Our strategy for detecting multiple diseases is supported by a goal-based evaluation structure. Evaluations that are based on goals determine whether or not goals have been achieved. In the multiple disease detection system, machine learning algorithms will extract data from two different sources. The first is brand-new information provided by people, while the second is a dataset that is obtained from Kaggle. The dataset is then split into two parts: a training set and a testing set. 25% of the data was used for testing, while the remaining 75% was used for training.

Graphic representation of our project design is drawn below:

****

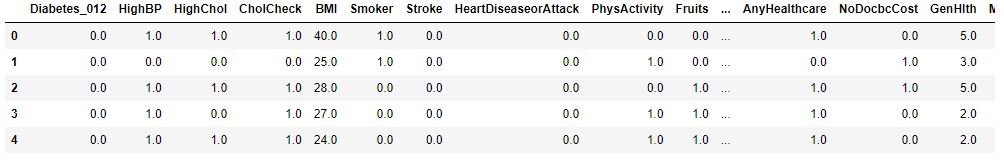
**Figure 3.1:** Approached credit card fraud detection system

**3.1.1 Experimental Dataset**

In our work on multiple disease prediction using machine learning, we used three distinct types of datasets. In order to build a perfect model that can accurately predict whether a person has these (heart, diabetes, or lung) problems or not, we utilize the dataset as an ideal one for the purpose of diagnosing illness.

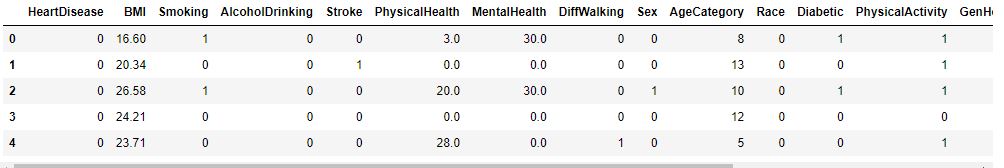
This project makes use of a Kaggle dataset that includes various amount of data. The amount of data for three different disease is given below.

1. Diabetes Disease: This data set contains 253681 data of patient where 213703 patient didn’t have any diabetes and the rest of them having diabetes.



**Figure 3.1.1.1:** Sample Data set for diabetes detection

1. Heart Disease: This data set contains 319794 data of patient where 292422 patient didn’t have any heart disease and the rest of them having diabetes.



**Figure 3.1.1.2:** Sample Data set for heart disease detection

1. Pneumonia Disease: This data set contains 11697 image data of patient where 3158 patient didn’t have any pneumonia problem and the rest of them having diabetes.

About the data collection:

1. Diabetes Disease: The data collection has a total of 17 features. 16of these characteristics are independent variables, while one is a dependent variable. The terms "independent set" and "dependent set" are mentioned here.
2. Heart Disease: The data collection has a total of 21 features. 20 of these characteristics are independent variables, while one is a dependent variable. The terms "independent set" and "dependent set" are mentioned here.

**3.2 Experiments for Pneumonia**

**3.2.1 Introduction**

In order to apply the most practical and acceptable model for the project, we have carried out the following experiments:

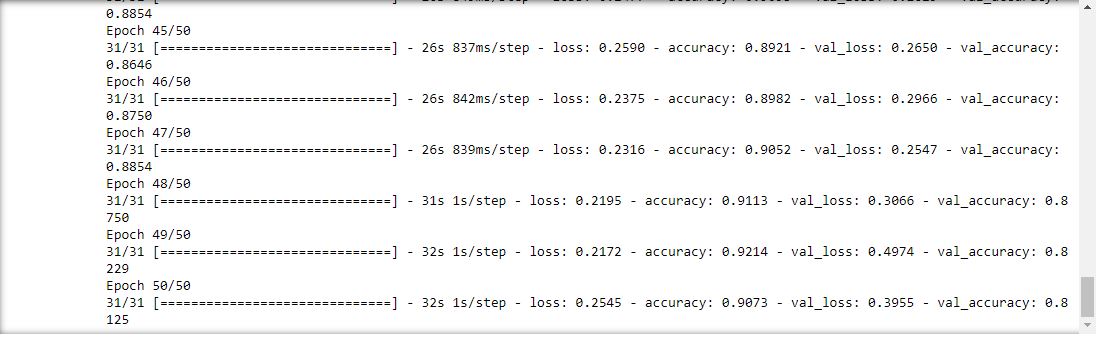
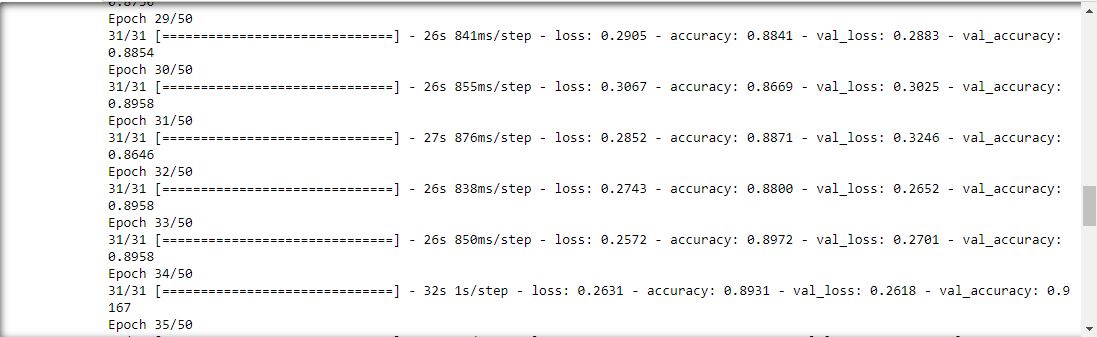
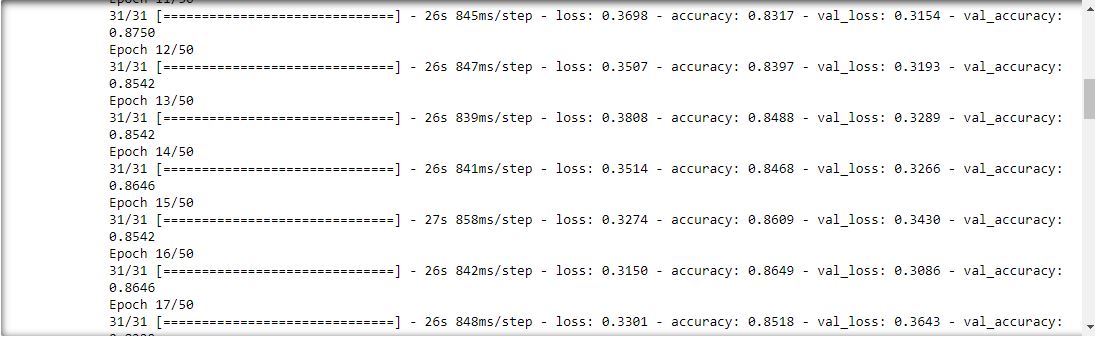
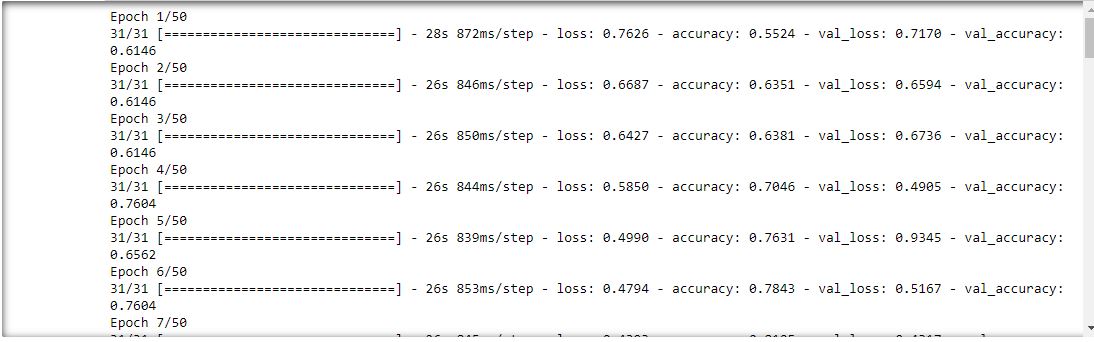
|  |  |  |  |
| --- | --- | --- | --- |
| Models | Activation Function (Input) | Activation Function (Hidden Layer) | Activation Function (Output) |
| Model-1 | ReLU | ReLU | ReLU |
| Model-2 | TanH | TanH | TanH |
| Model-3 | ReLU | ReLU | Softmax |

**3.2.2 Model-1 Training with Rectified Linear Unit**

Rectified Linear Unit (ReLU) activation function in neural networks is extensively used for machine learning.

The ReLU function returns the input value if it is larger than zero and zero otherwise using the formula f(x) = max(0, x). To put it another way, ReLU leaves all positive integers alone while reducing all negative values to zero.

The ReLU activation function has quickly become well-known in the deep learning community due to its computational effectiveness and practical success. Additionally, utilizing sigmoid or tanh activation functions, it helps resolve the vanishing gradient problem that can occur in exceedingly deep neural networks.

****

**Figure: Training with Model-1**

**3.2.3 Model-2 Training with Tangent Function**

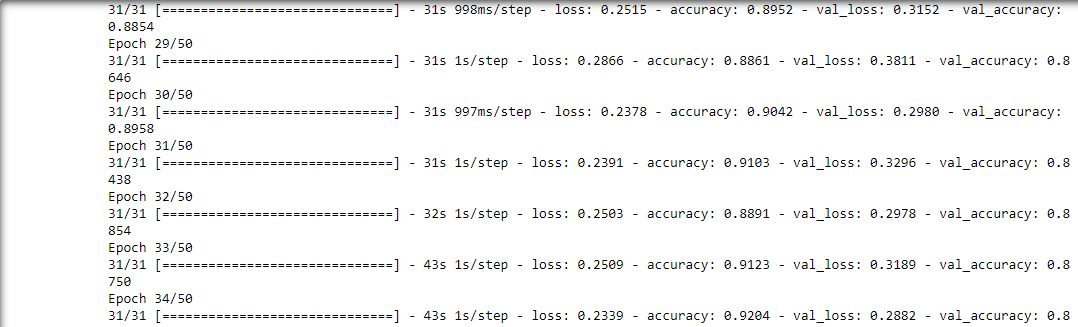
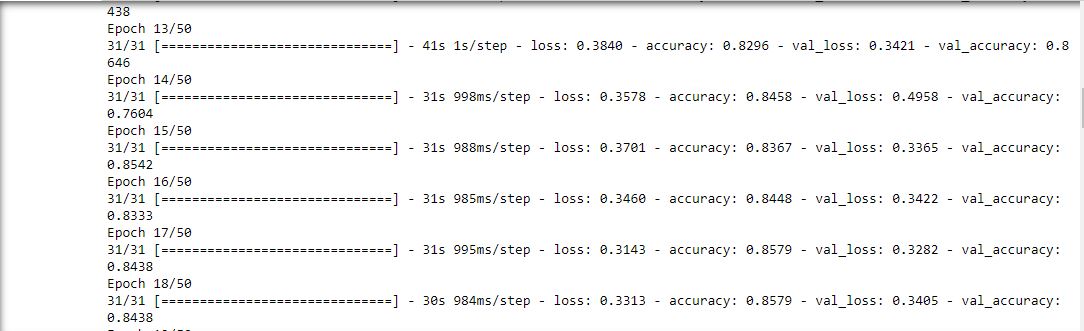
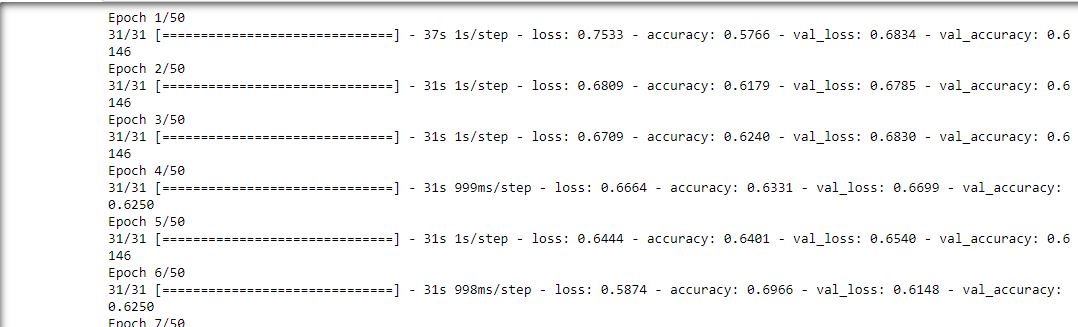
The hyperbolic tangent function (abbreviated "tanh") is a mathematical function that applies to a number between -1 and 1. (Ex - E-X) / (Ex + E-X) = tanh(x) is how it is defined.

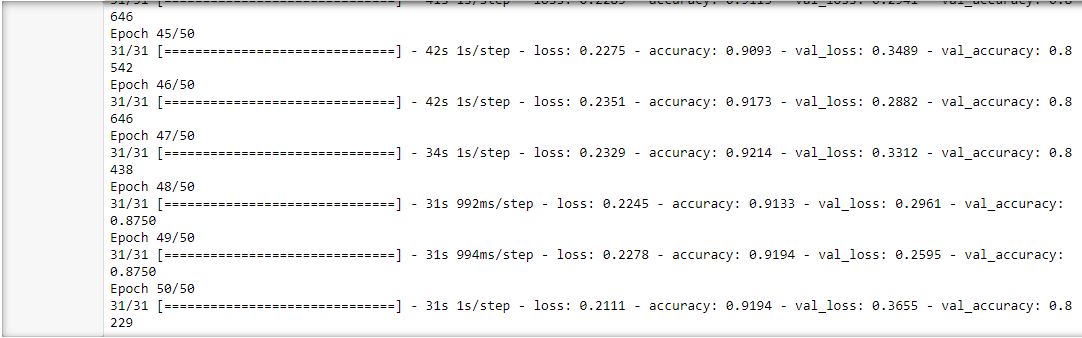
where x is the variable being evaluated and e is a mathematical constant that roughly equals 2.71828.

A curve with a maximum at x=0, a minimum at -1 as x tends towards infinity, and an approach to 1 as x tends towards +infinity is used to graphically illustrate the hyperbolic tangent function.

Because of the y-axis symmetry of the function, tanh(-x) = -tanh(x).

3.2.4 Model-1 Training with Rectified Linear Unit and Softmax





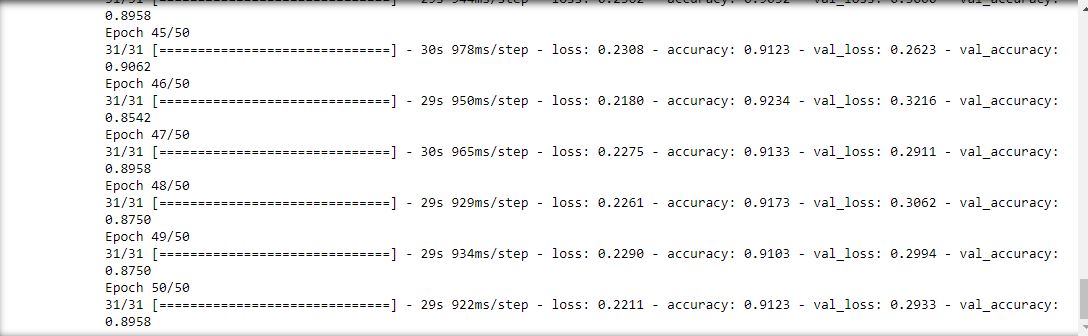
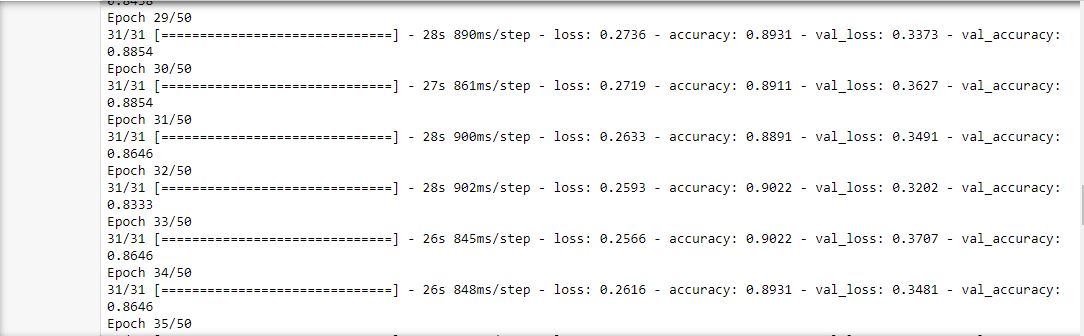
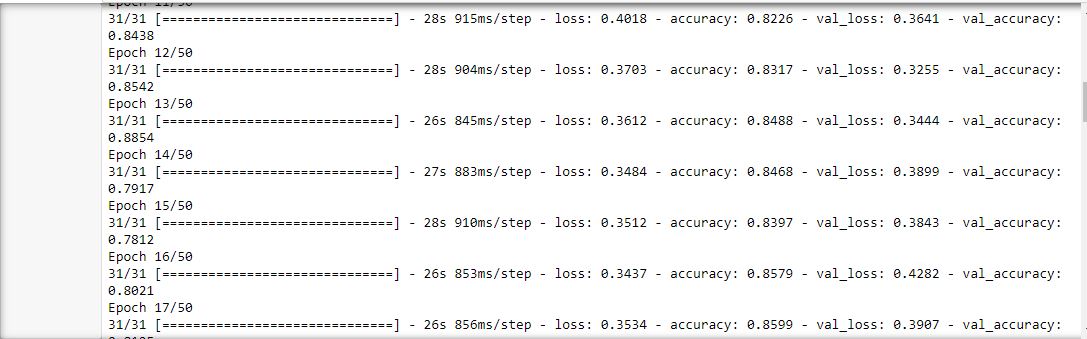
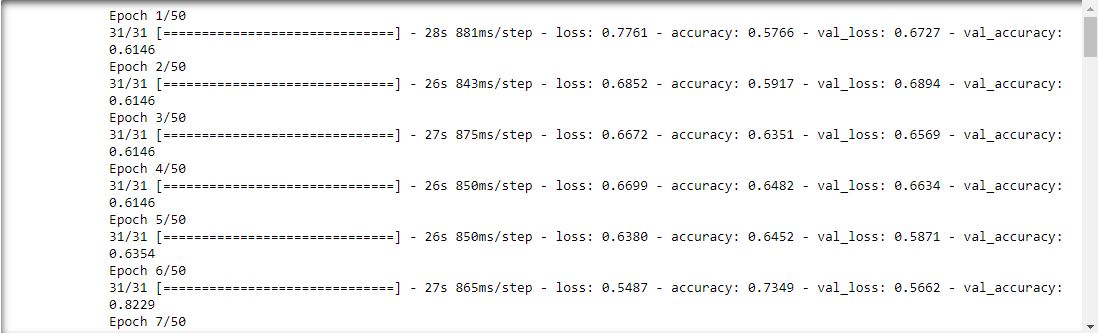
**3.2.4 Model-1 Training with Rectified Linear Unit and Softmax**

Neural networks typically use the softmax activation function to create output probabilities for classification tasks. An input vector of arbitrary real-valued scores is used to compute a probability distribution over a set of classes.

The softmax function normalizes the output vector of exponentiated scores after multiplying each input value by an exponent so that they add to one. When given an N-dimensional input vector x, the softmax function is defined as follows:

Softmax(x\_i) = exp(x\_i) / sum(exp(x\_j) for i = 1,..., N.

where the denominator's total is computed across all j items, and j is a component of the input vector.



**3.3 Experiments for Diabetes**

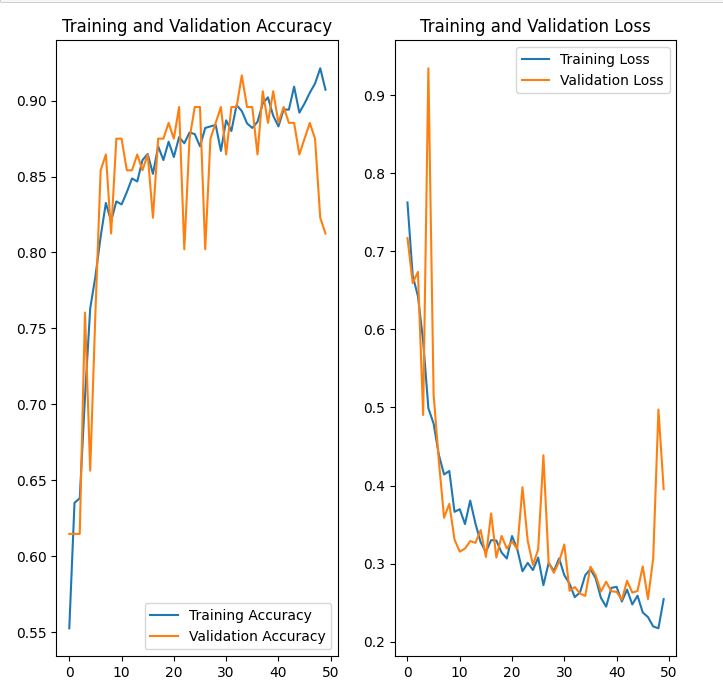
**3.3.1 Introduction**

**3.2 Results**

We utilize CNN machine learning technique for Pneumonia and three distinct machine learning techniques, divided into two groups: classification and regression. They are included here, along with the results of each algorithm’s classification report.

**3.2.1 Outcome of Model-1 (Pneumonia)**

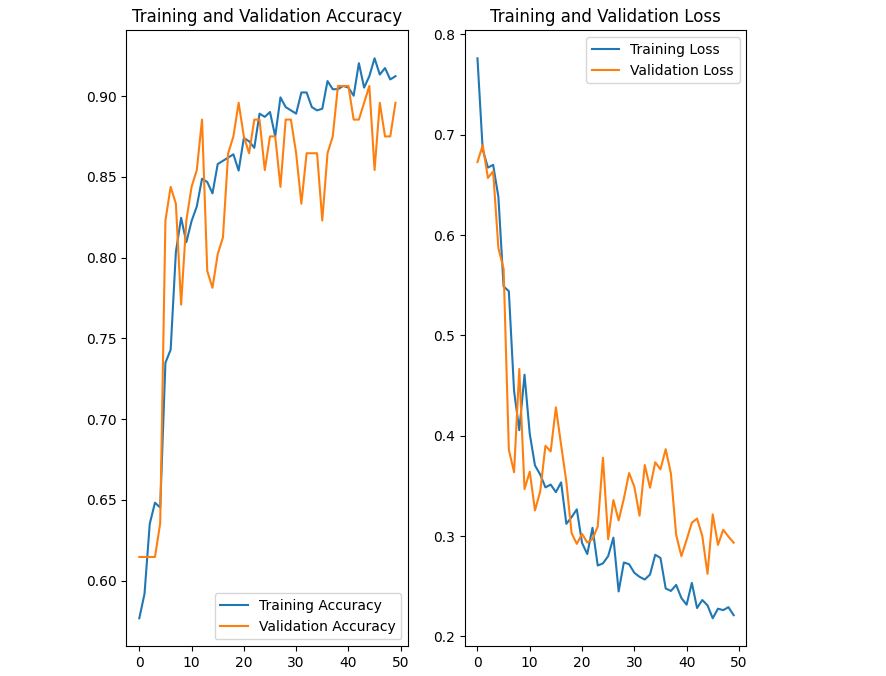
Convolutional layers are used by the CNN to identify patterns in incoming data and make predictions. These data are transmitted to the fully connected levels. The discrepancy between expected and actual output during network training affects neuron weights. The accuracy and precision of the network are improved with each repeat by lowering the error margin. Here, the blue line stands in for training, and the orange line for validation. The first model demonstrates a gradual increase in training and validation accuracy. In training and validation loss, the reverse situation is seen. This indicates that the model gradually develops throughout the epochs. The accuracy during training is almost 1, while the accuracy during validation is between 0.85 and 0.95.



**Figure 3.2.1:** Performance Graph of Model-1

**3.2.2 Outcome of Model-2 (Pneumonia)**

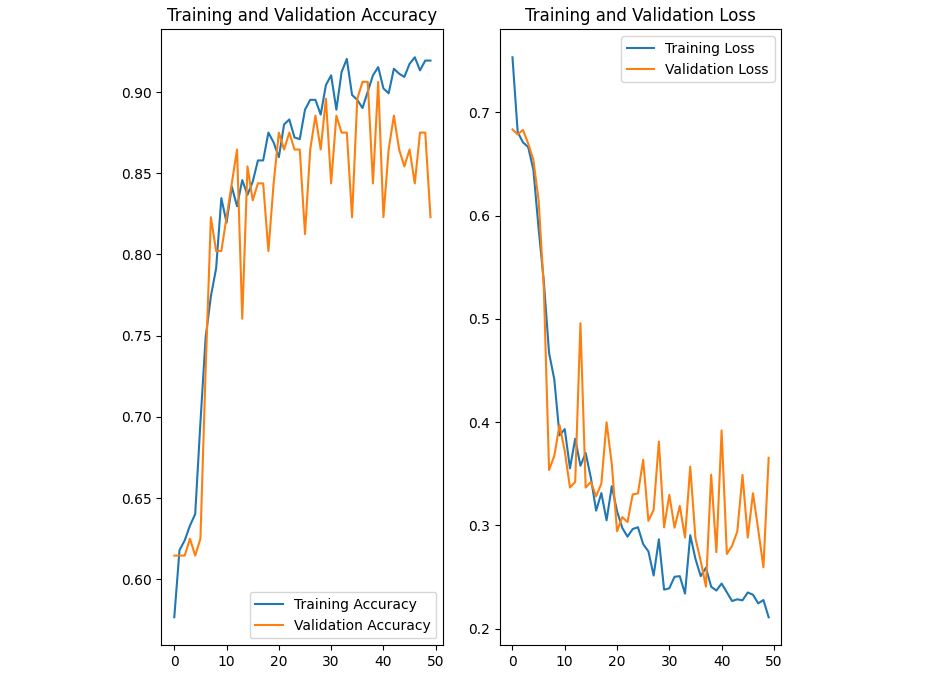
Overfit networks struggle with novel input and memorize training data. To get around this, a validation set is constructed using some of the training data to test how well the network performs on unidentified data. Instead of forcing the network to memorize the training set, the data structure can be tailored to. Validation accuracy ranges from 0.86 to 0.91, and training accuracy is almost 0.87. Since the lines are almost completely overlapping, it just recalls the training data. The model is hence overfit. Additionally, the performance is subpar due to the low precision.



**Figure 3.2.2:** Performance Graph of Model-2

**3.2.3 Outcome of Model-3 (Pneumonia)**

This makes it obvious that the model has a remarkably low level of accuracy. Between 0.82 and 0.92 is the range when the training set's accuracy is deemed to be at its highest. Additionally, the maximum accuracy for the validation set is between 0.82 and 0.92. The model's performance falls short of expectations due to its extremely low accuracy, despite the fact that the training and validation loss is quite little in this instance.



**Figure 3.2.3:** Performance Graph of Model-3

**3.3 Discussion**

The hypothesis we offered in our study is really well-fitting. Because our data set is well-trained, and we applied a new machine learning technique, we were able to detect fraud transactions with a better rate of accuracy. The data set we used is severely unbalanced, with just 7200 fraudulent transactions out of 594,643 total. That is why we attempted to use an algorithm that would deliver a greater rate of accuracy.

We gain 99.4 percent overall accuracy using the decision tree technique, but only 75 percent precision when detecting fraudulent transactions. We get about 85 percent fraud transaction precision using the KNN algorithm, 94 percent fraud transaction precision using SVM, 43 percent fraud transaction precision using naive bayes, and 87 percent fraud transaction precision using logistic regression.

As a result, decision trees, KNN, Naive Bayes, and logistic regression algorithms did not work any better in the face of a highly skewed dataset. However, when dealing with a severely unbalanced credit card fraud data set, the SVM algorithm performs substantially better. As a result, we can put the SVM algorithm in charge of detecting credit card fraud in the real world.

A comparative result of our proposed model is mentioned here:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **AUC** |
| Decision Tree | 99.4% | 75% | 77% | 76% | 1796 | 88.28% |
| KNN | 99.4% | 83% | 59% | 69% | 1796 | 79.65% |
| SVM | 99.3% | 94% | 47% | 63% | 1796 | 73.39% |
| Naïve Bayes | 98.5% | 43% | 78% | 55% | 1796 | 88.36% |
| Logistic Regression | 99.3% | 87% | 54% | 67% | 2226 | 76.94% |

**Table 3.3:** A comparative result of our proposed model using machine learning algorithm

A comparison of the results of our suggested model using five different classification and regression algorithms. Among these algorithms, SVM has a greater precision rate and a higher accuracy rate. As a result, our model can determine if a new transaction is fraudulent and send a message to the card issuer.

**Chapter 4**

**Related Work**

In this section, we go through some of the earlier work that has been done on Fraud Detection.

Machine learning techniques play a critical role in a variety of efficient data processing domains, one of which is the detection of card fraud. Several ways for detecting fraud were recommended in prior studies, including supervised methods, unsupervised methods, and a hybrid strategy; this necessitates knowing some technology involved in recognizing credit card fraud, as well as a deeper understanding of the sorts of card fraud. Many other solutions were proposed and tested. The majority of them will be discussed in the following paragraphs.

Yashvi Jain, Namrata Tiwari, Shripriya Dubey, and Sarika Jain researched SVM, ANN, Hidden Markov Model, Bayesian Networks, KNN, Fuzzy Logic system and Decision Trees in 2019. In their research, they discovered that the algorithms k-nearest neighbor, decision trees, and SVM provide a medium level of accuracy. Fuzzy Logic and Logistic Regression have the lowest accuracy of all the methods. Neural networks, naive bayes, fuzzy systems, and KNN all have a high detention rate. The LogisticRegression, SVM, decision trees have a high detection rate at the middle level. Both ANN and Nave Bayesian Networks outperform each other across all parameters. It is highly expensive to train these people. Each algorithm has a unique set of parameters. They give great results with one type of dataset, but they produce horrible results with another type of dataset.[1]

The work of Navanushu Khare and Saad Yunus Sait on decision trees, random forests, support vector machines (SVM), and logistic regression was published in 2018. They had to deal with a dataset that was considerably biased. Accuracy, sensitivity, specificity, and precision are used to evaluate performance. Overall accuracy of Logistic Regression is 97.7%, Decision Trees is 95.5 percent, Random Forest is 98.6%, and SVM classifier is 97.5 percent, according the statistics. Among the various ways, they determined that the Random Forest approach has the highest accuracy and is the best algorithm for detecting fraud. They also revealed that the support vector machine (SVM) approach has a data imbalance problem and performs poorly in detecting credit card fraud.[1]

Developments in E-Commerce and Communication technology, according to Altab Althar Taha and Sareef Jameel Malbery, have created credit card usage a far more frequent mode of payment, and transactional fraud is on the rise. Researchers used the high optical support vector machine, which combines Bayesian-based high energy optimization with parameter adjustment of the light gradient boosting machine (LightGBM). They employed a real-world public dataset that included both fraudulent and non-fraudulent transactions in their technique. Their suggested approach exceeded other strategies in terms of accuracy when compared to other techniques. The suggested system achieves a 98.40 percent accuracy, a 92.88 percent area under the receiver operating characteristics curve (AUC), a 97.34 percent precision, and a 56.95 percent F1-score.[4]

Debachudamani Prusti and Santhnu Kumar Rath designed an application employing machine learning methods such as decision trees, K-nearest neighbor (KNN), multilayer perceptron (MLP), and support vector machine to determine the accuracy in fraud detection (SVM). They proposed a model that included decision trees, support vector machines, and k-nearest neighbor algorithms. They used two web-based protocols for successful data exchange across several diverse platforms: simple object access protocol (SOAP) and representational state transfer (REST). They used the accuracy metric to compare the results of machine learning algorithms. The support vector machine (SVM) outscored other algorithms by 81.63 percent, while the hybrid system they presented had an accuracy of 82.58 percent.[4]

Because fraud detection must be very flexible in order to track the continuous growth of fraud over time and the appearance of unknown anomalies, H.Tran and K.P.Tran used anomaly detection techniques for credit card fraud detection. As data-driven techniques, they recommended a T2 control chart and a one-class support vector machine OCVM with optimum kernel parameter selection. The approach was tested utilizing a huge real-time data set of online e-commerce transactions from European credit card holders, which contained a total of 284807 non-fraud transactions. Simulators were also used to create fraudulent transactions, with 284000 transactions used for training and 200 fraud and non-fraud transactions used for testing. To examine the outcomes obtained by the techniques, they used accuracy, F1-score, Recall (DR), FPR, and Precision matrices. The experimental results show that OCVM surpasses the T2 flowchart with accuracy of 96.6 percent, FPR of 8.5 percent, and F-score of 100 percent. The two proposed techniques, on the other hand, have shown that they can accurately detect credit card fraud with a low false rate.[5]

**Chapter 5**

**Future Work**

Based on the results of the above research, we were unable to achieve our aim of 100 percent accuracy in fraud detection, indicating that there is still space for improvement. For detecting fraud, machine learning techniques are applied, although the results are not accurate.

There are certain approaches that can be used to correctly identify credit card fraud. As a result, implementing deep learning algorithms will improve the accuracy of detecting credit card fraud.

Another alternative for improvement is to expand the amount of the dataset; this would enhance the precision of the used algorithm. These will provide considerably more precise results. As an outcome, more data will almost certainly enhance the model's effectiveness in fraud prevention while also reducing the false positive rate. However this needs support (by giving more personal information about users) from the banks themselves.

**Chapter 6**

**Conclusion**

As credit cards carry not only money but also intellectual property, credit card fraud is a serious crime. To lessen the severity of this criminal crime. Machine learning can be a useful tool for detecting fraudulent transactions. We applied several classification and regression algorithms in our credit card fraud detection project. These algorithms give us a considerably greater level of accuracy in detecting fraudulent transactions. Because the SVM algorithm reacts significantly better in response to our severely imbalanced dataset, it fared the best among these algorithms.

The SVM (Support Vector Machine) technique gives a 99.3 percent accuracy rate and a 94 percent precision rate in the fraud class. In the actual world, that is a far higher accuracy rate for detecting fraud transactions. When a new transaction occurs, it will be processed using our proposed machine learning model, which will use the SVM method to determine if the transaction is fraudulent or not.

We may not be able to detect fraud transactions with 100% accuracy, but we can use these datasets in conjunction with real-world data. After that, we can reach 100% accuracy by integrating some machine learning techniques. Despite the fact that credit card data is confidential, different banks will not provide their genuine credit card user data set.

**Bibliography**

[1] R. Tyagi, R. Ranjan, and S. Priya, “Credit Card Fraud Detection Using Machine Learning Algorithms,” *Proc. 5th Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2021*, no. Iciccs, pp. 334–341, 2021, doi: 10.1109/I-SMAC52330.2021.9640822.

[2] S P Maniraj, Aditya Saini, Shadab Ahmed, and Swarna Deep Sarkar, “Credit Card Fraud Detection using Machine Learning and Data Science,” *Int. J. Eng. Res.*, vol. 08, no. 09, pp. 110–115, 2019, doi: 10.17577/ijertv8is090031.

[3] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, “An efficient credit card fraud detection model based on machine learning methods,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 3414–3424, 2020.

[4] A. RB and S. K. KR, “Credit card fraud detection using artificial neural network,” *Glob. Transitions Proc.*, vol. 2, no. 1, pp. 35–41, 2021, doi: 10.1016/j.gltp.2021.01.006.

[5] H. Najadat, O. Altiti, A. A. Aqouleh, and M. Younes, “Credit Card Fraud Detection Based on Machine and Deep Learning,” *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, no. Section IX, pp. 204–208, 2020, doi: 10.1109/ICICS49469.2020.239524.