DETECTING CERVICAL CANCER USING LSTM, DBN, ANN ALGORITHMS

LAVANYA GUPTA

1 VELLORE IINSTITUTE OF TECHNOLOGY, CHENNAI

2Department of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600001, India

3Computer Science specialization in robotics and artificial intelligence, Vellore Institute of Technology, Chennai, India

author: Lavanya Gupta (e-mail: lavgupta0801@gmail.com).

ABSTRACT This study explains the application of Artificial Neural Networks (ANN), Deep Belief networks (DBN) and long short -term memory networks (LSTM) for the detection of cervical cancer through detailed study, analysis and research. we evaluate and compare the accuracy of these advanced machine learning techniques against the traditional methods. Our findings indicate the significant improvements in the accuracy of technology and efficiency of different algorithms like ANN, DBN, LSTM models achieving the accuracy of 97.76,93.29,97.12 percent respectively. Here we will be discussing about the methodologies employed for data processing, model training, and performance evaluation, highlighting the key insights and comparative advantages of each approach. This study underlines the potential of neural networks in enhancing cervical cancer detection and outlines future avenues for clinical integration and further advancements.

INDEX TERMS Cervical cancer, Artificial neural networks (ANN), Deep belief networks (DBN), Long short-term memory networks (LSTM), machine learning, Feature extraction, Predictive model, Precision, f1 score

1. INTRODUCTION

Cervical Cancer is one of the most significant global health burden, this is mainly profound in the countries which do not have access to regular screening and early detection methods. According to WHO the cervical cancer remains as one of the most affecting Cancer among women with annual deaths ranging to 311000[1]. Recently the advancement that is involved in the field of Artificial Intelligence led to the prior detection cervical cancer in the early stages using the algorithms like ANN, DBN, SVM, LSTM, RF, LR [2].This paper aims at improving and proper detection of cervical cancer using ANN,LSTM,DBN models leveraging recent advancements in machine learning algorithms, Here we will be highlighting the methodologies, findings, and implication for clinical practice eliminating the problems of integration, providing comprehensive review and comparison, relevance and translatability and cost effectiveness.

2. LITERTURE REVIEW

The research paper uses the machine learning concepts like Random Forest (RF), Logistic Regression (LR), SMOTE preprocessing was used to resolve the data imbalance by producing synthetic samples for minority class. t-SNE feature extraction was used to capture complex structures in the data. SVM consistently outperforms RF and Logistic Regression when combined with SMOTE preprocessing and t-SNE feature extraction, demonstrating improved accuracy, precision, and recall rates, attaining the accuracy of 92.60 percent, precision of 0.91 and recall of 0.90. The software used was python.[1]

The work presents the working of a decision Tree classification algorithm which demonstrate the importance of feature selection for the prior level prediction of cervical cancer for improving the accuracy, sensitivity and specificity the missing values of the dataset were eliminated by the under and oversampling models called the SMOTETomek for eliminating the data imbalance showcasing the accuracy of 98 percent, sensitivity 0f 100 percent and specificity of 97 percent.[2]

The research paper deals with the prediction of cervical Cancer in females using the ensemble classifier and various machine algorithms the proposed ensemble based model classifies various textual dataset based on the hospital’s records of various patients with better prediction and accuracy [3]

The research paper related to ensembled based model was again proposed, the proposed work shows the result of various machine learning algorithms like Shapley additive values (SHAP), local interpretable model agnostic explanation (LIME), random forest and ELI5 have been effectively utilized., the one way ANOVA, mutual information and Pearson’s correlation techniques are used for feature selection SMOTE was employed in the algorithm for the class imbalance removal. The final stacked machine learning model obtained an accuracy, precision, recall, F1-score, area under curve (AUC) and average precision of 98%, 97%, 99%, 98%, 100% and 100%, respectively [4]

The research paper provides a predictive model for cervical cancer using a Kaggle based dataset which contains no missing data attaining a accuracy of 96.8 percent using the decision tree algorithm along with the lowest RMSE (root mean square error) and RAE (relative absolute error) of 0.15 and 41.6. [5]

In this paper [6] the model employed was Naïve Bayes

Classification algorithm based on the idea of conditional probability which means the probability of occurrence of an event which is dependent on various other events. The algorithm was proposed as it is easy to implement and also the use of a large dataset achieving an accuracy of 87 percent

This paper [7] provides a comparative study of the prediction of Cervical Cancer using various machine learning algorithms like KNN, RF, LR, Naïve bayes, SVM with accuracy of 57, 59, 61, 78, 65 respectively using a group of 858 female patients aged between 13 to 84 from a Venezuelan inpatient clinic. The data being used for previously processed, subjected to principal component analysis, and feature selection was based on the removal of highly correlated and near zero variance predictors. The data used here is divided into an 80:20 train-test split ratio to evaluate the model performance of data outside the training set.

The research paper [8] focusses on the on risk factors, prevention, and management of cervical cancer. The paper uses the algorithms like KNN (k -nearest neighbor), Logistic Regression and Support Vector Machine (SVM) attaining the accuracy of 94.2 percent using SVM.

In this research paper [9] aims at separating the healthy cells with unhealthy cells with an accuracy of 1.00 classifiers were applied to the dataset .it also is used in distinguishing the cervix and CIN with healthy tissue with 0.95 accuracy using SVM machine. The dataset being split into training and testing groups in the ratio of 6:4 Of the 210 included patients, 166 found healthy, 22 were having eroded cervix , and 31 were diagnosed of cervical intraepithelial neoplasia (CIN). The support vector machine (SVM) algorithm was employed for the classification of normal and abnormal cervical tissue governed by 11 characteristics features.

In this paper [10] aims at finding an efficient machine learning based classifier and models for the prior and early detection of the cervical cancer. A Random forest algorithm provided the best classification method with accuracy of 98.33 and KNN with accuracy of 98.56 for schiller dataset .Both method uses the sine and logarithmic functions .various feature searching algorithms were employed.

In this research paper [11] which aims at prediction of cervical cancer using a proposed model named as Boruta Algorithm which aims at using the statistical procedures and followed by the series of continuous execution of the employed algorithm RF’s to evaluate the significant of authentic predictors to arbitrary , as such edge factors attaining an precision is 0.912, recall is 0.891, *F*1 score is 0.798, and support is 0.768.

**III. MATERIALS AND METHODS**

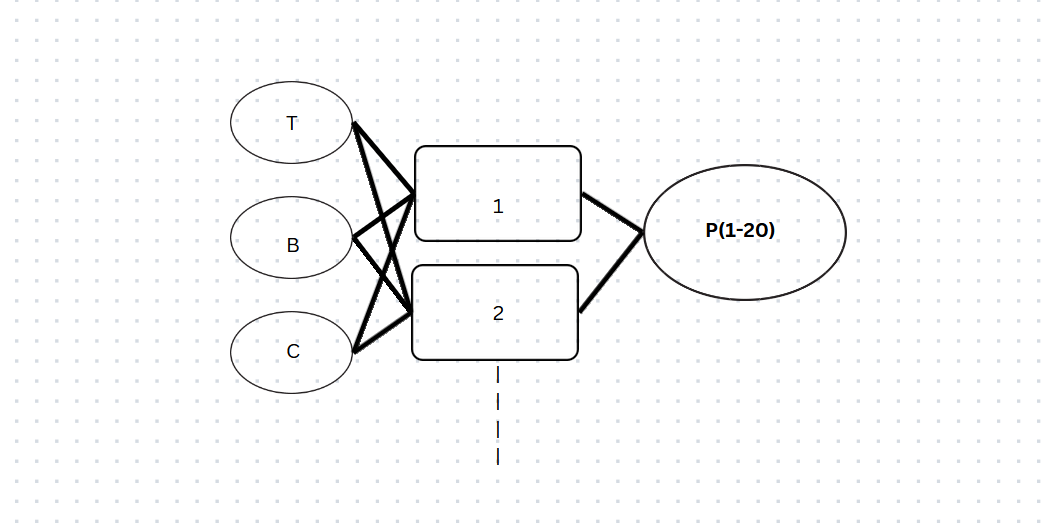
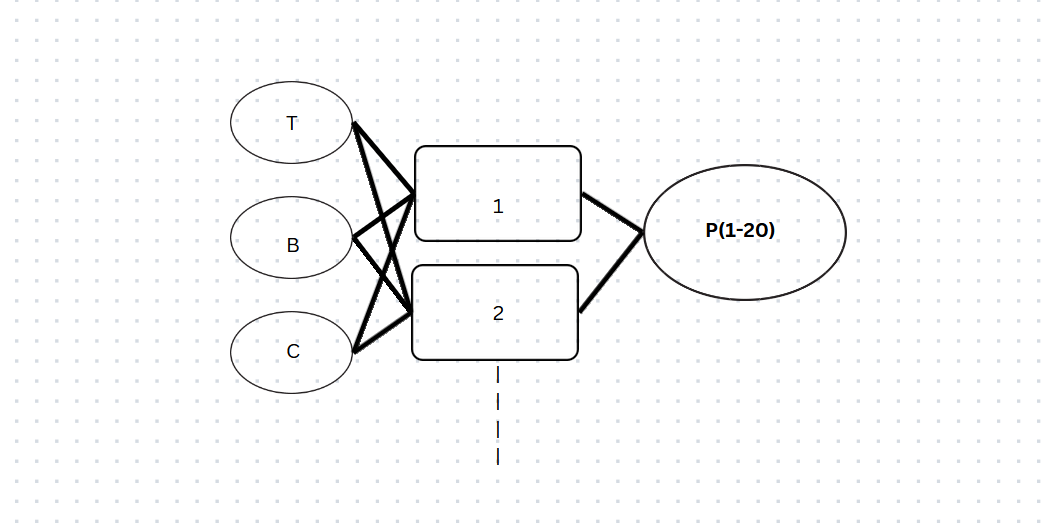
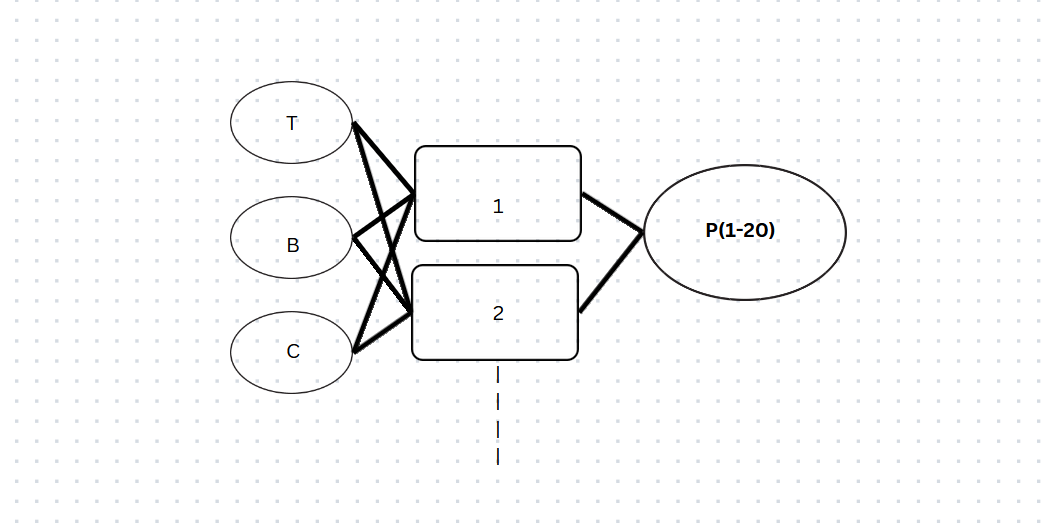
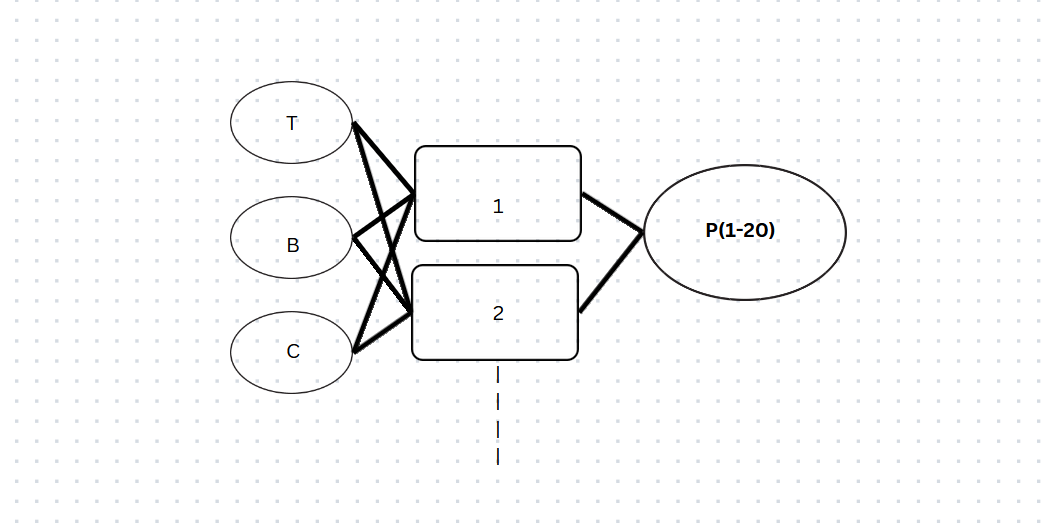
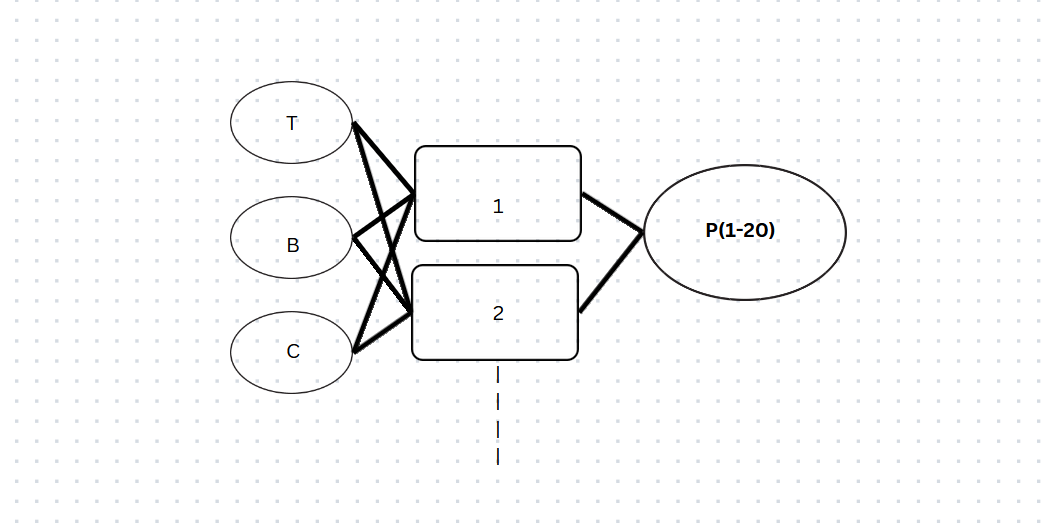
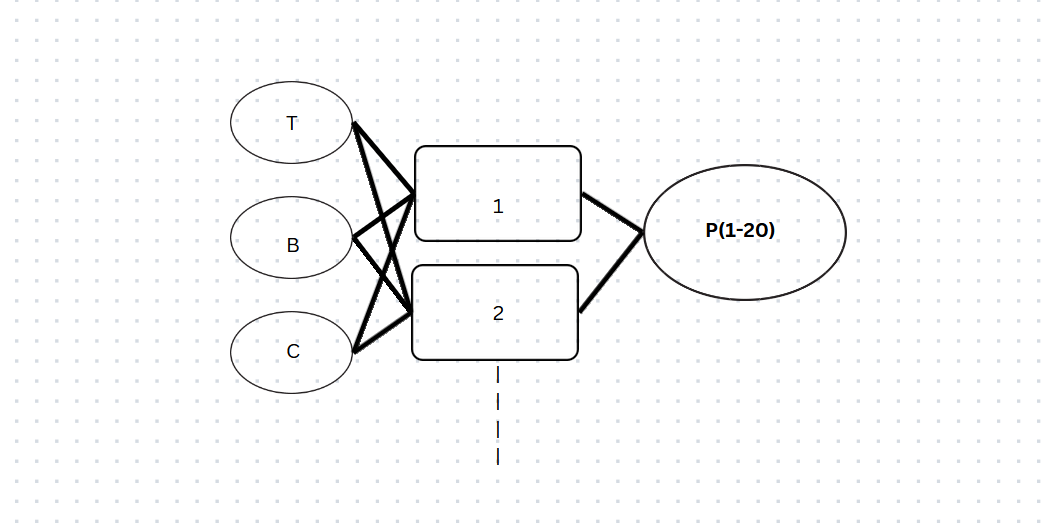
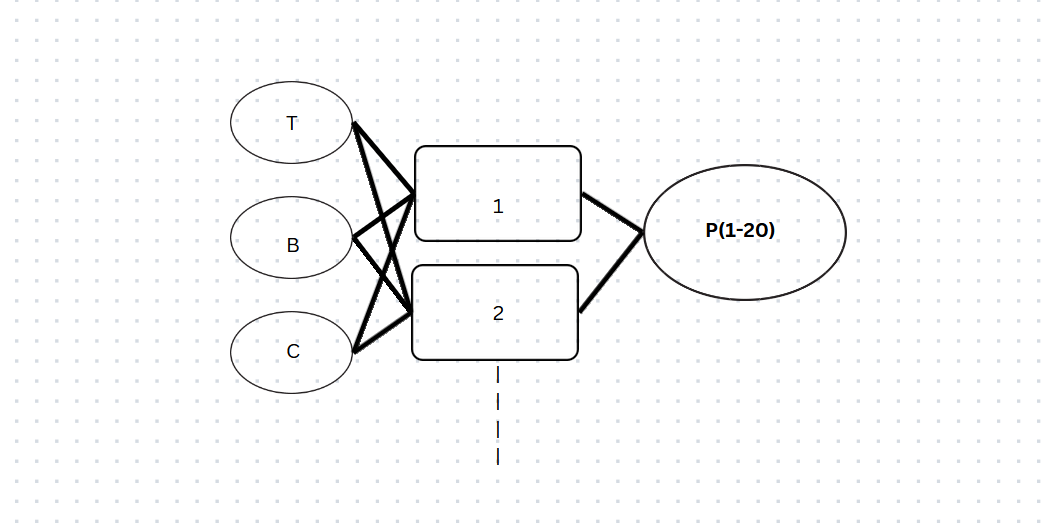
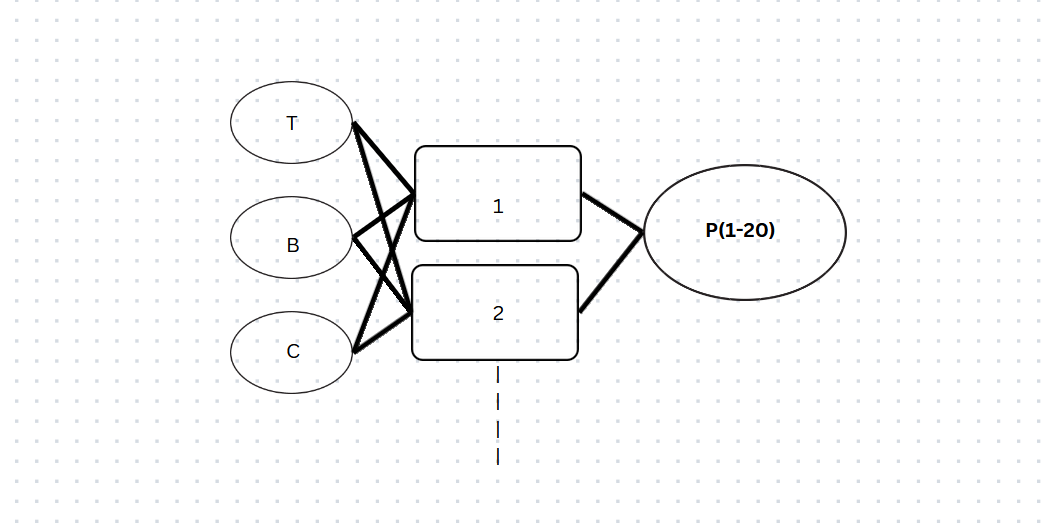
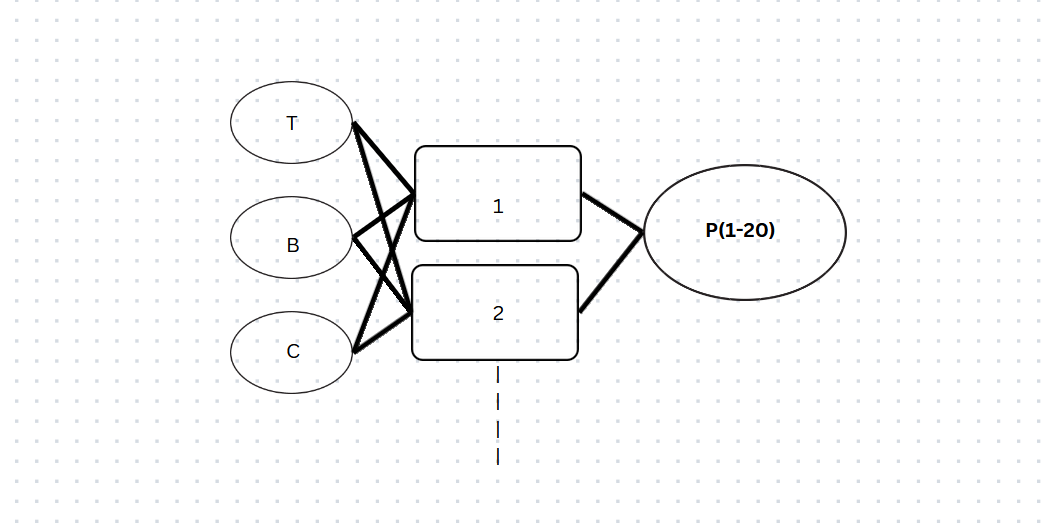
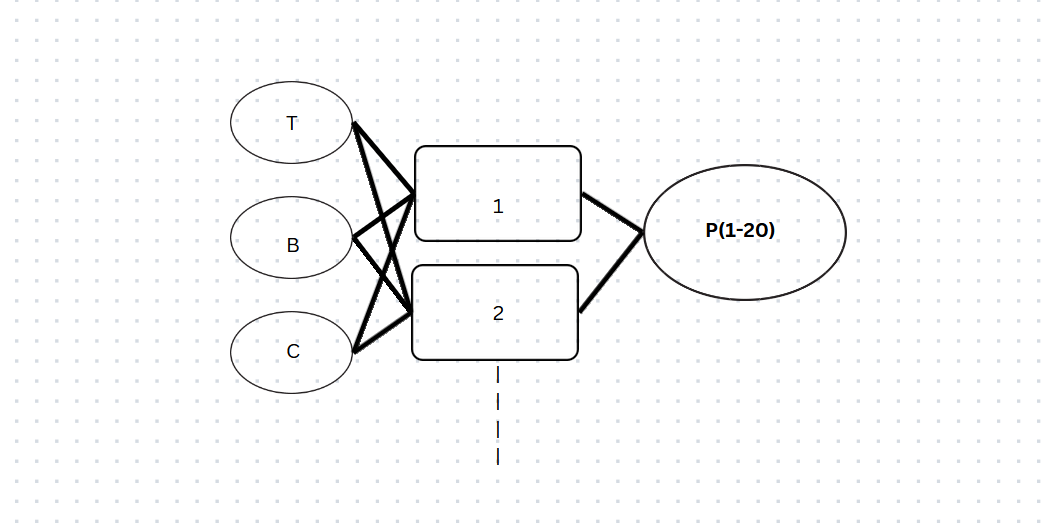
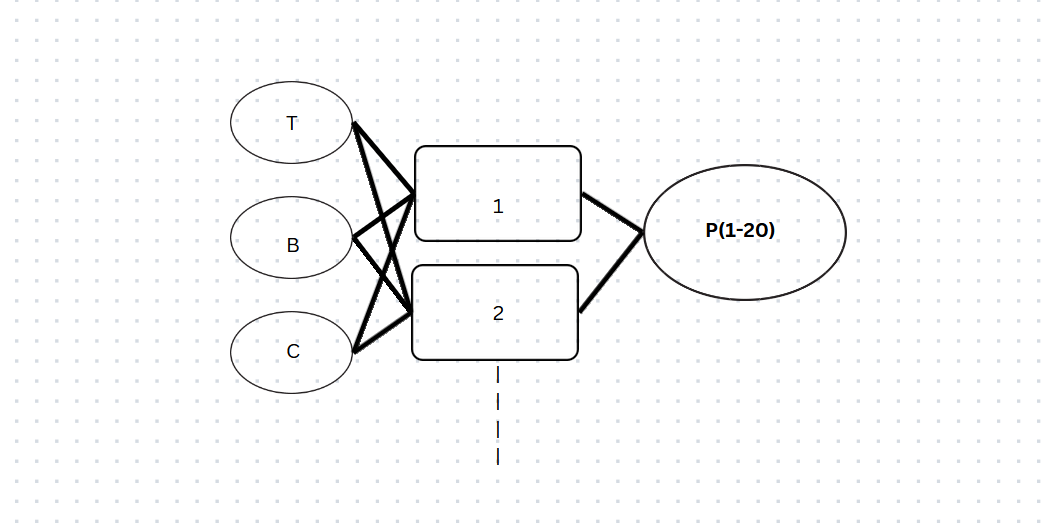
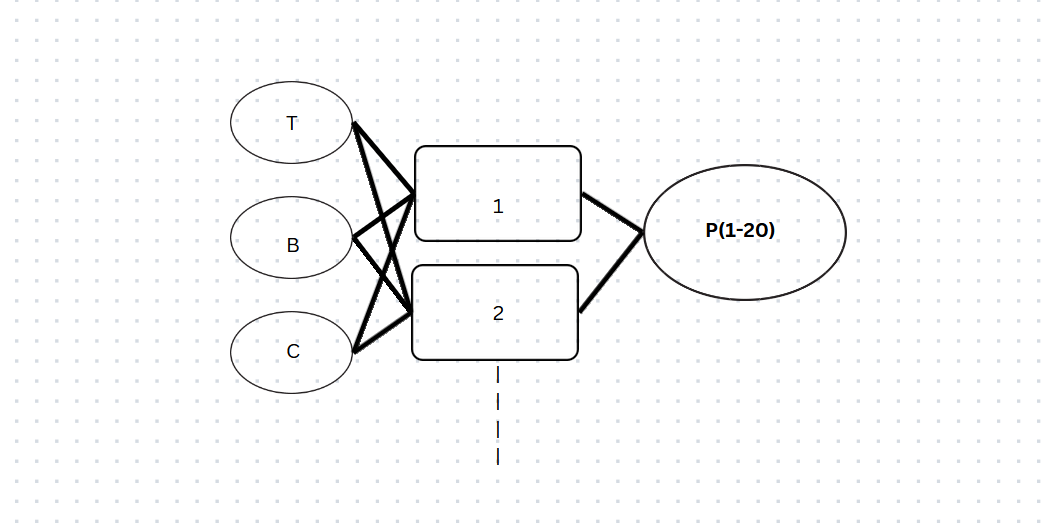
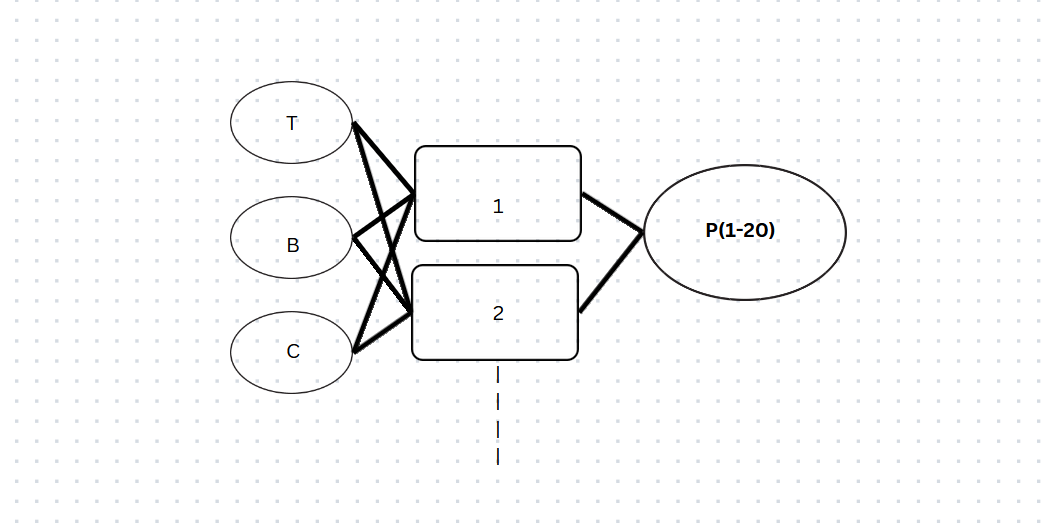
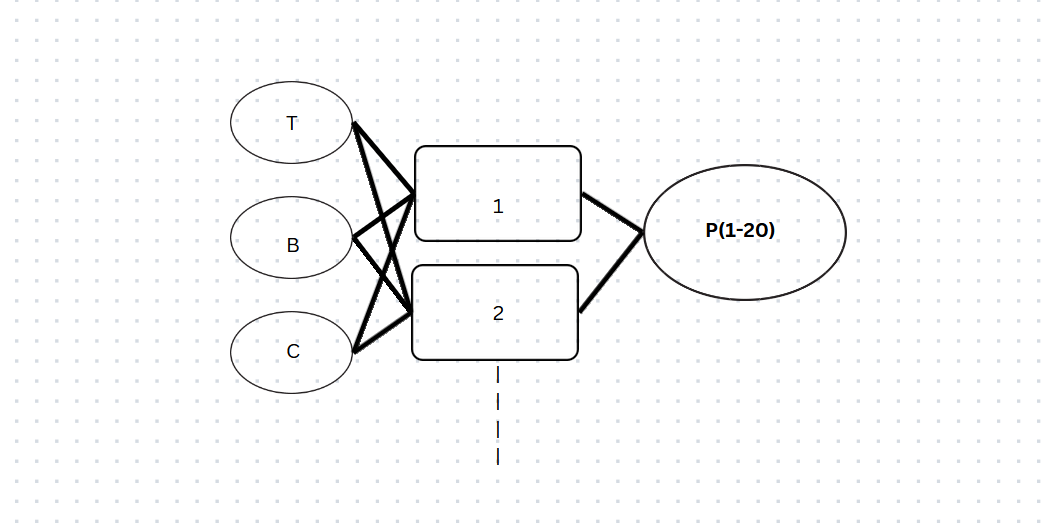
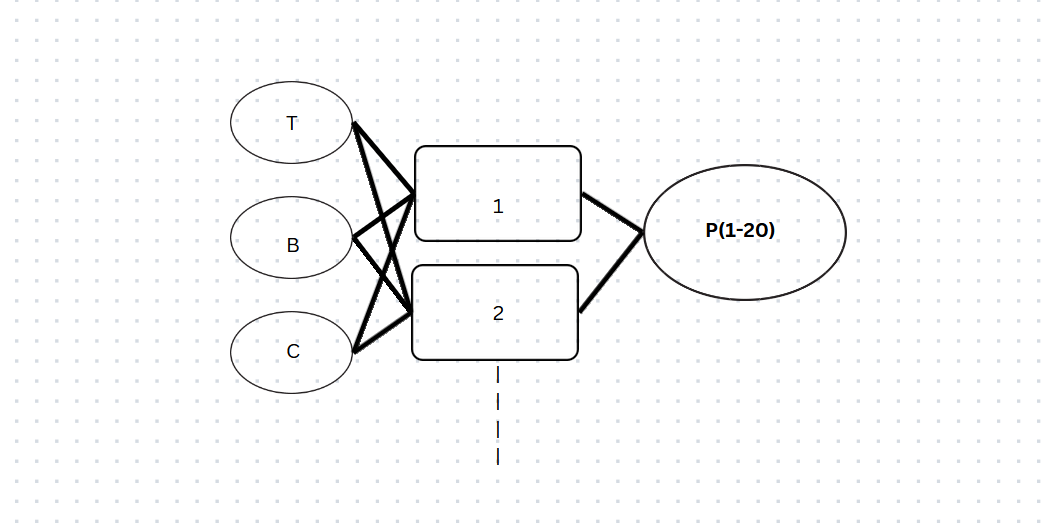
The proposed model is working on the prediction of cervical Cancer in women in early stages using different machine learning algorithms like ANN (artificial neural networks), LSTM (long-short term memory), DBN (deep belief network)

Modelled on ANN architecture

The description about the method and algorithms used are given below:

1)Artificial Neural Networks (ANN): ANN is the most fundamental deep learning algorithm which is inspired by the biological algorithms. The model consists of various layers having various interconnected nodes which help in studying the given input data and prepare the output predictions.

The various parts of ANN are:

              **INPUT LAYERS|HIDDEN LAYERS|OUTPUT LAYERS**

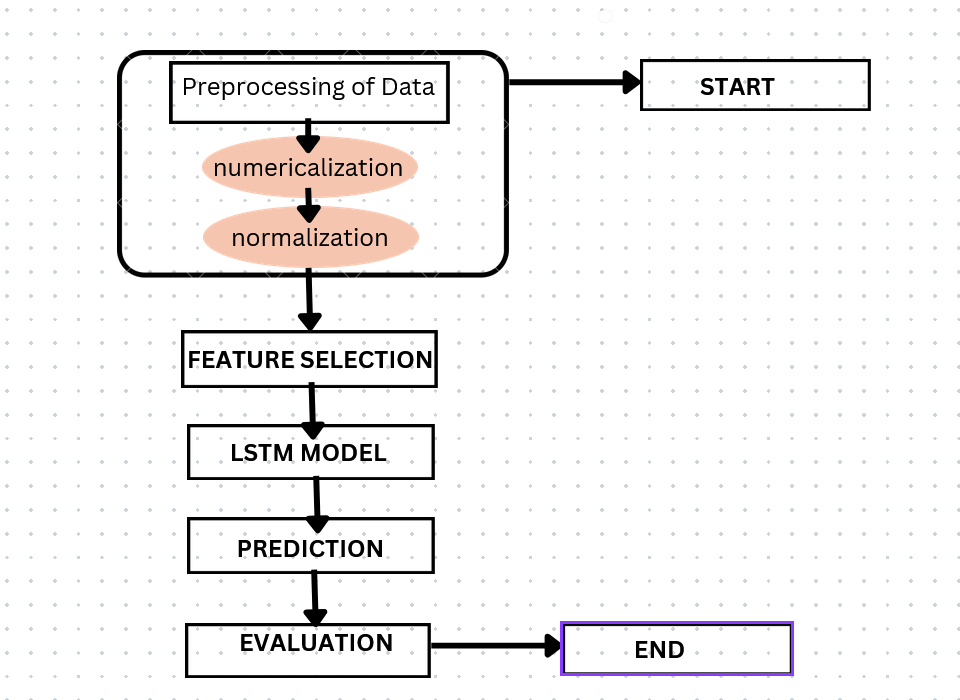
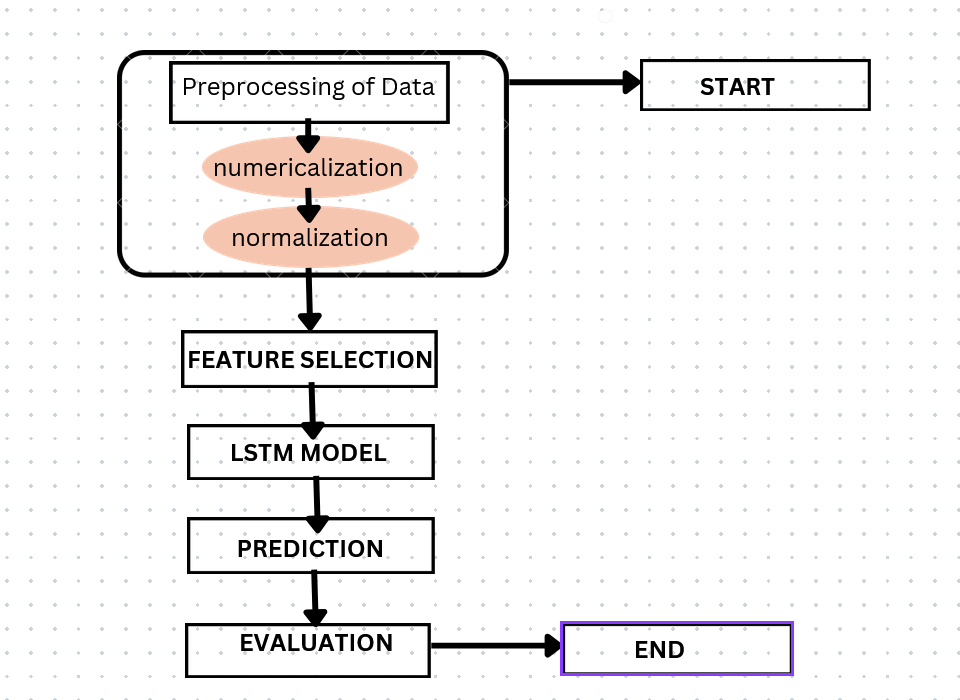
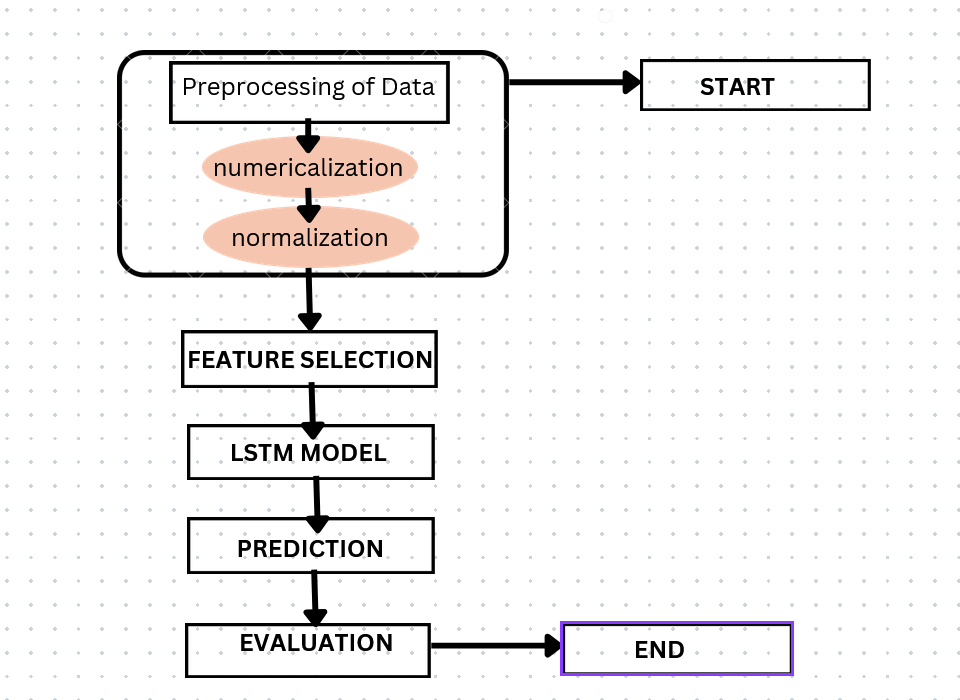
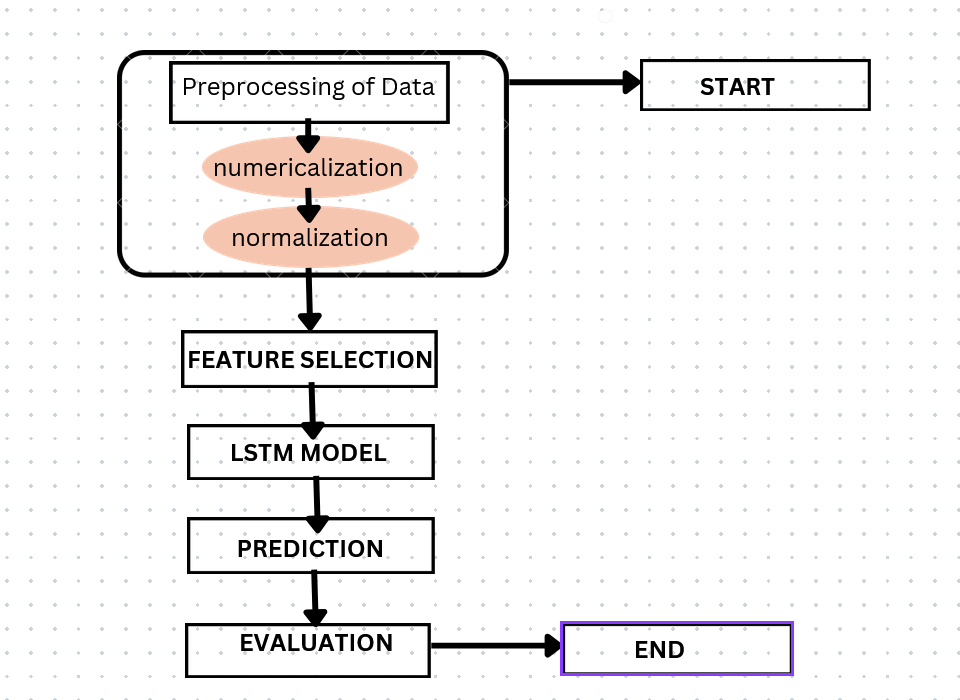
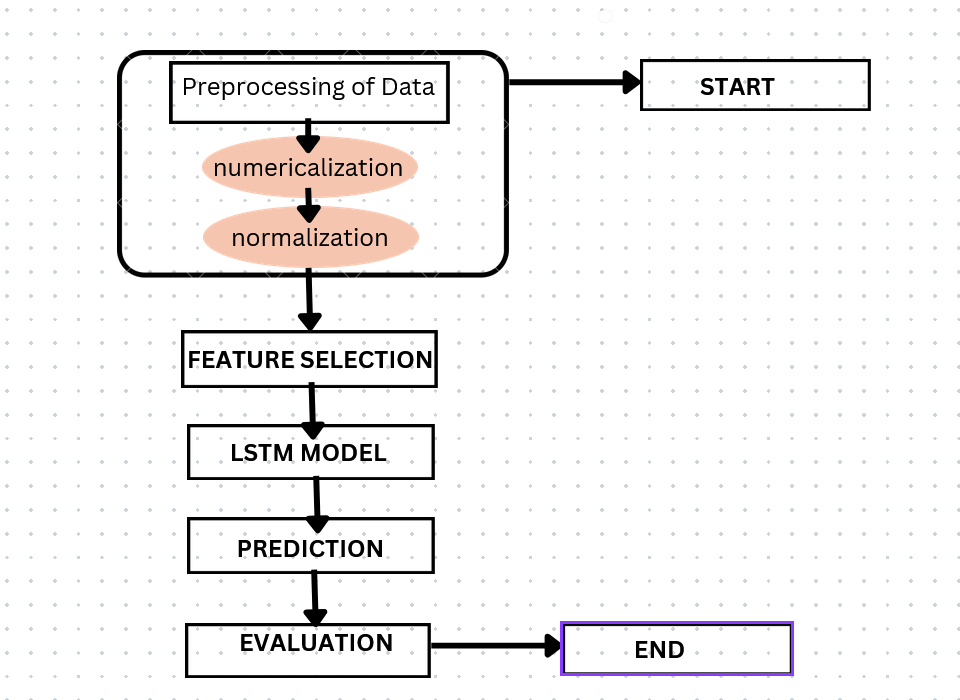
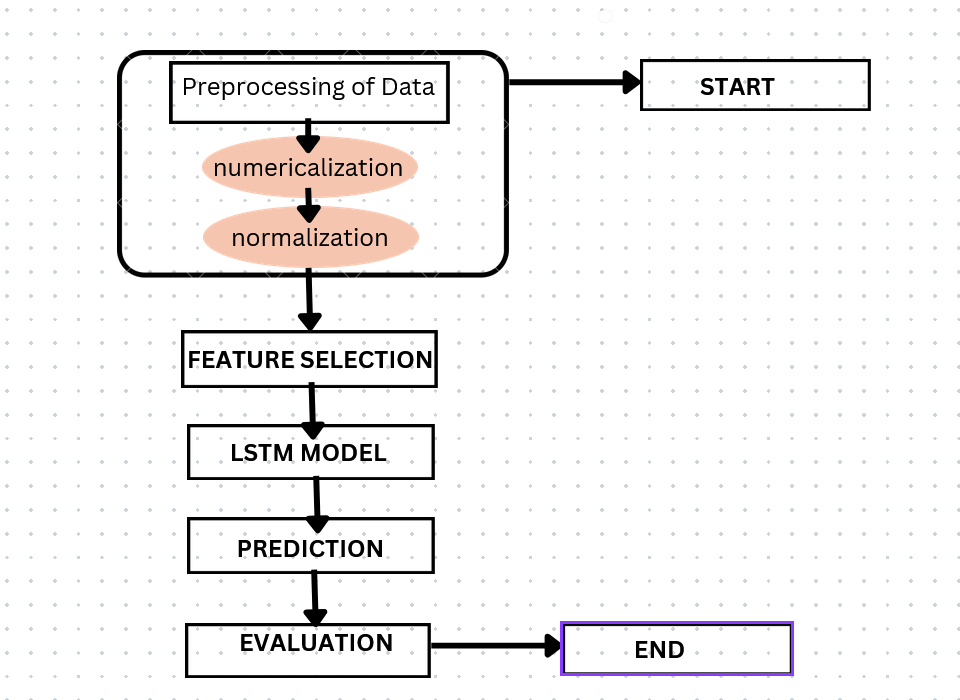
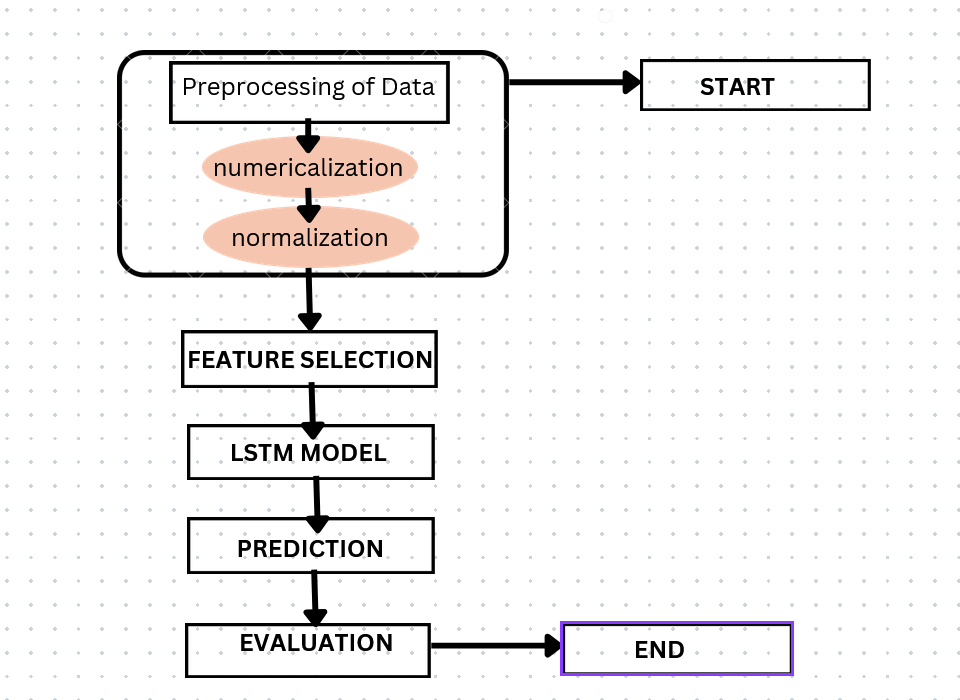
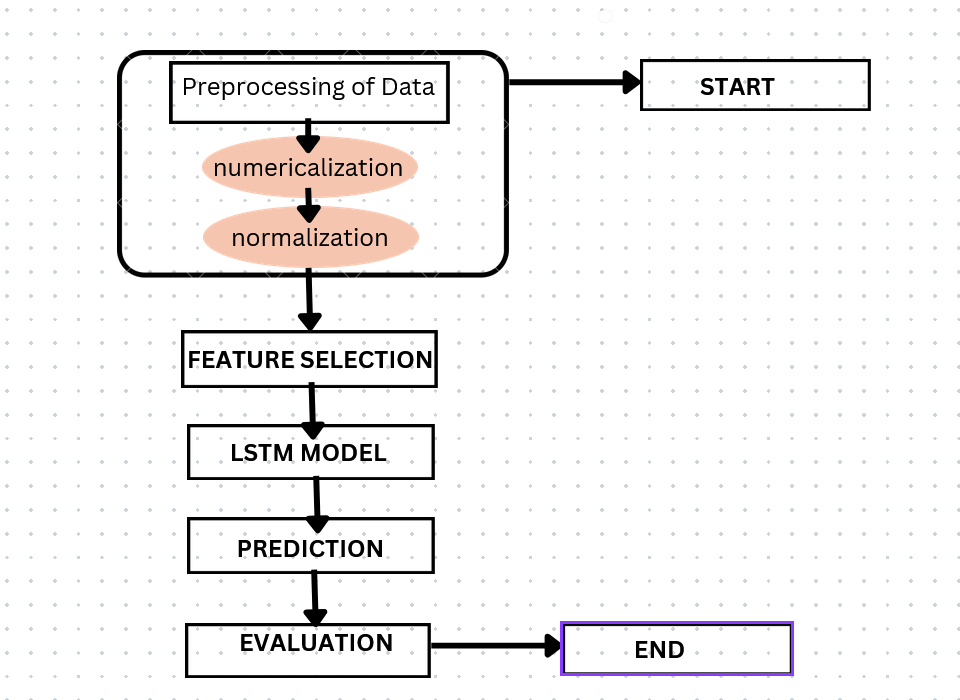
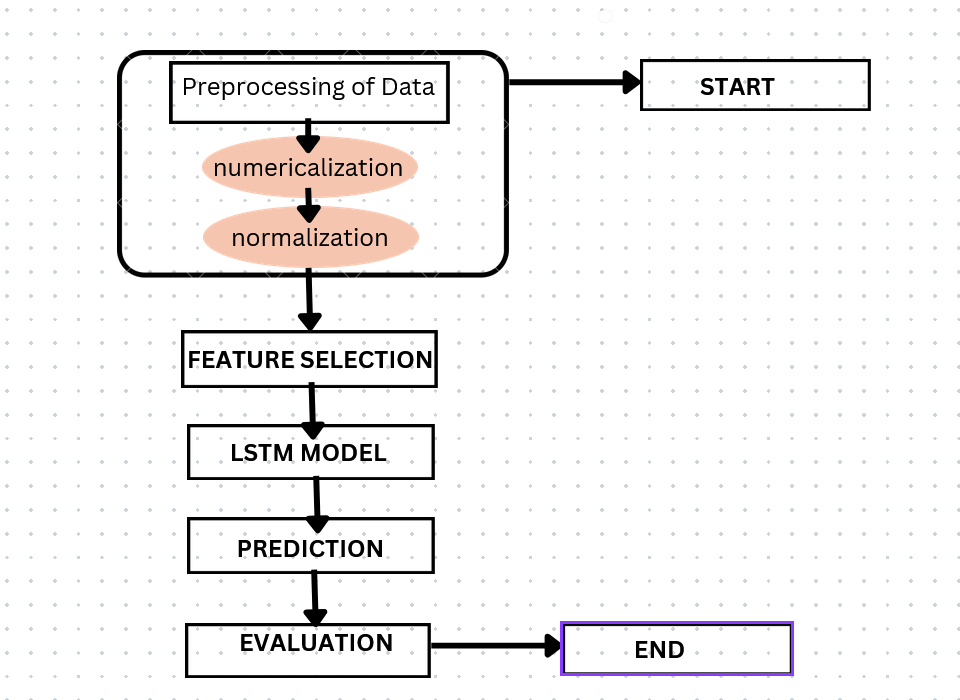
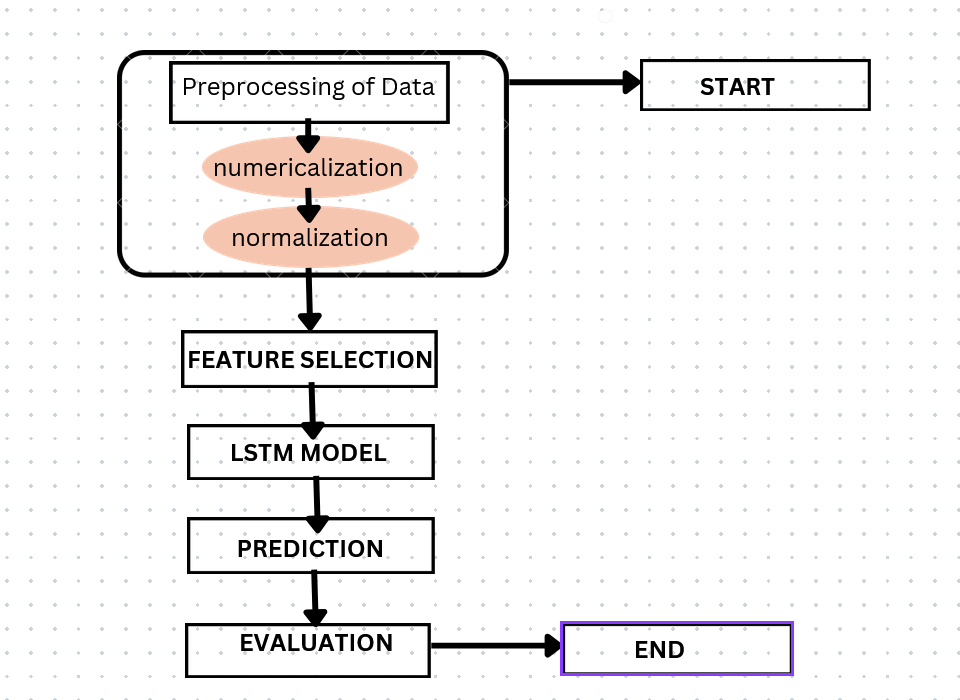
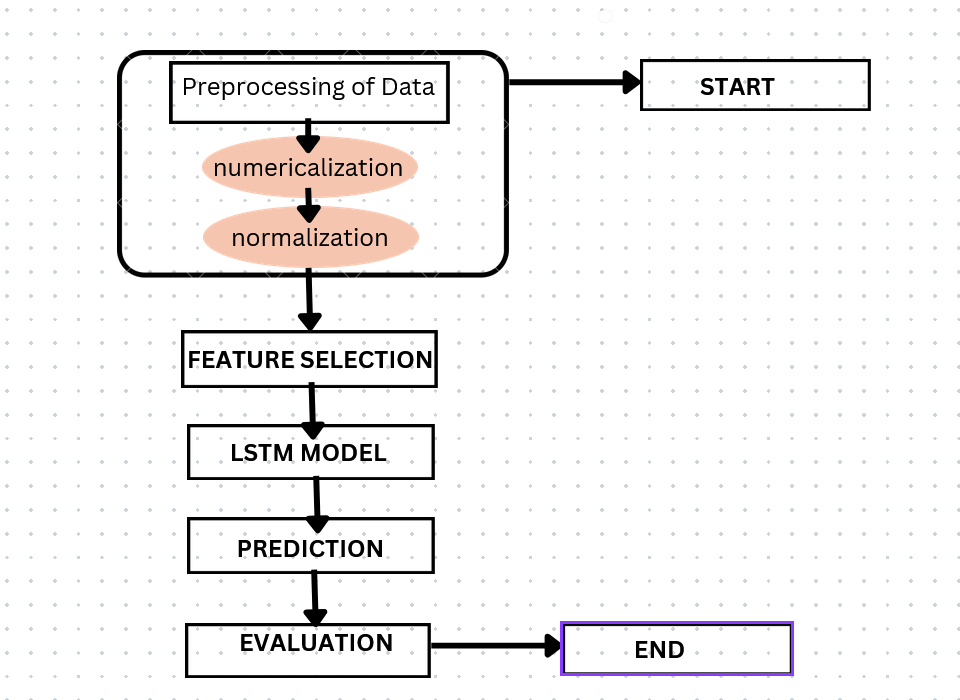
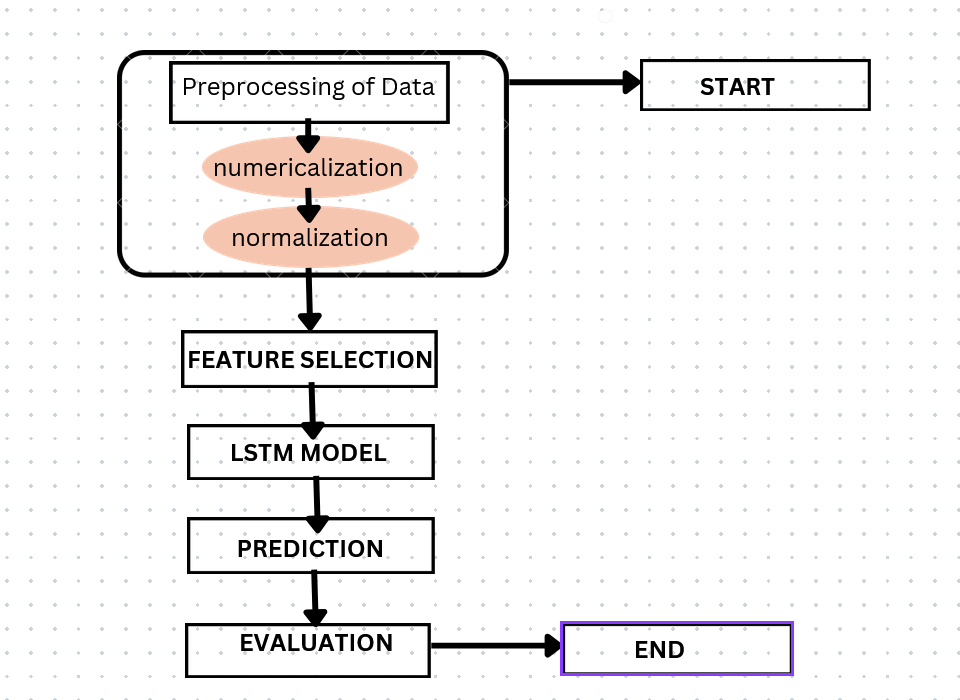
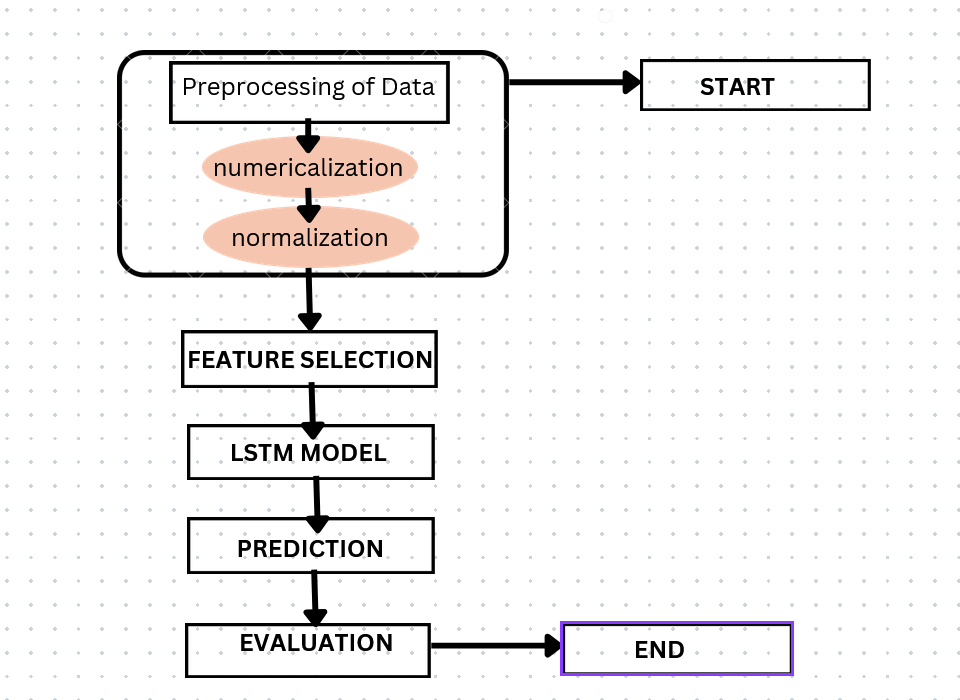
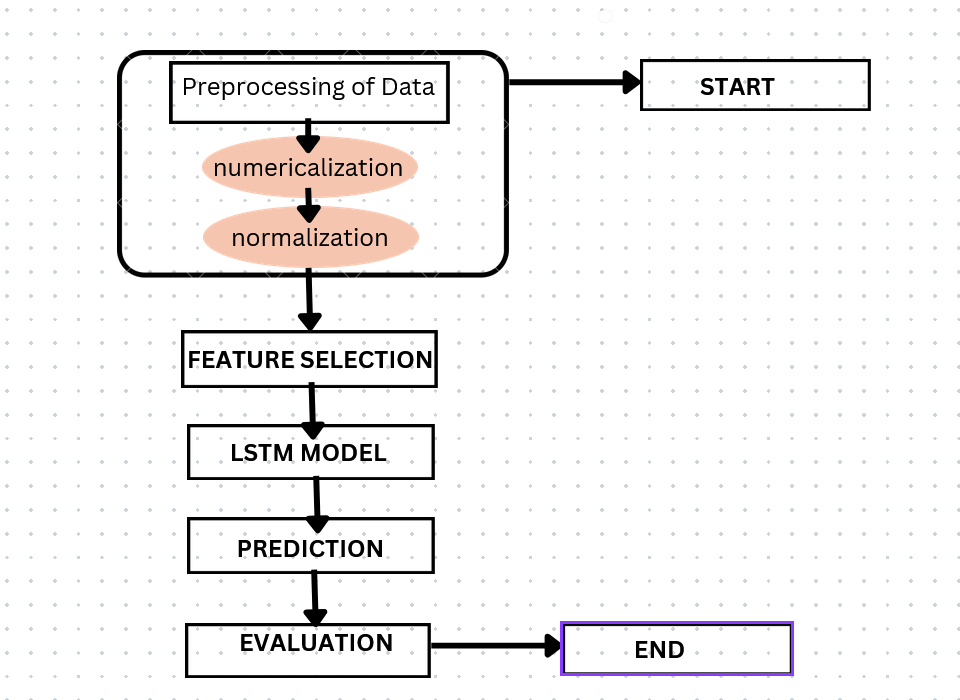
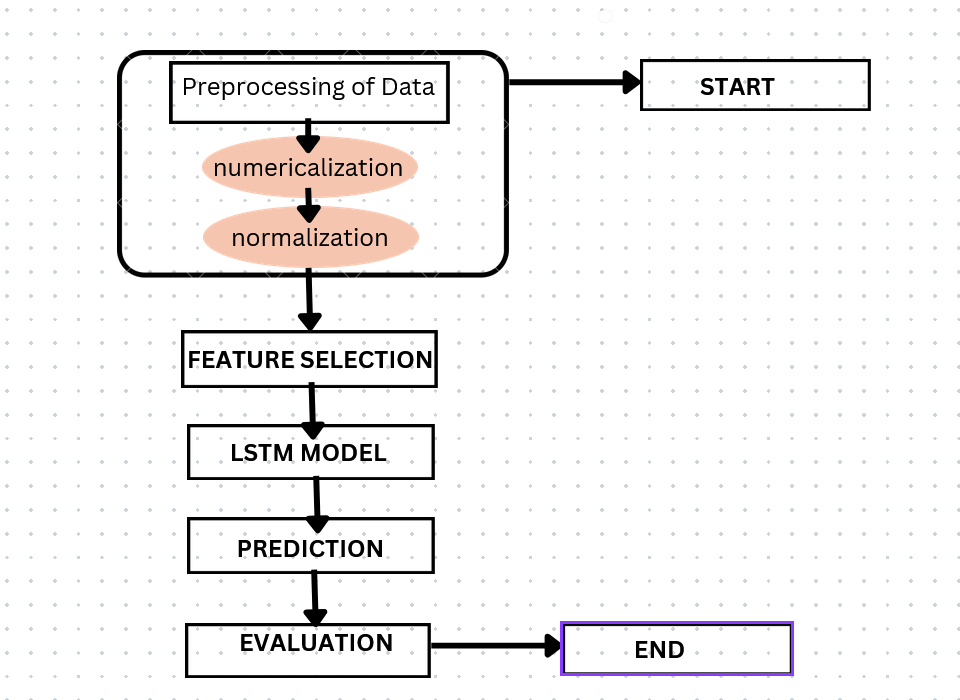
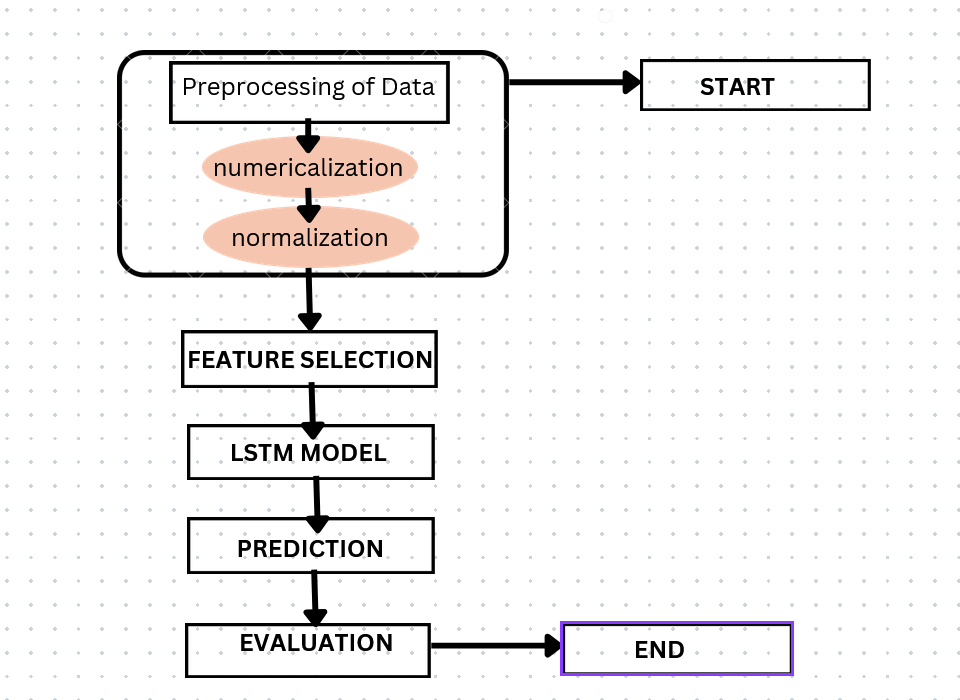
INPUT Layers: These are used for the purpose of feature selection.

HIDDEN layers: comprise of many layers performs a weighted sum of inputs followed by a function

OUTPUT layers: generates the output as 0 or 1 for cervical Cancer

TRAINING: Uses backpropagation and gradient descent to optimize weights and biases, minimizing binary cross-entropy loss.

2)Long-Short Term Memory (LSTM)- LSTM is a model designed for the sequential data based on the recurrent neural network (RNN) as it maintains a long- term data dependencies. It is mainly used for time series because of its ability to remember the data for a long period of time.

**Fig** -**The LSTM MODEL ILLUSTRATION**

INPUT layers: takes input from dataset in the format (batch size, time steps, features).

LSTM layers: unit that contains sequence retaining the memory.

OUTPUT layers: generates the binary function using sigmoid activation function.

TRAINING: focusses on the sequence of features selection.

3)Deep Belief Network (DBN): they are a class of deep neural networks comprising of a large number of hidden units Here we are implementing it using the ANN architecture.

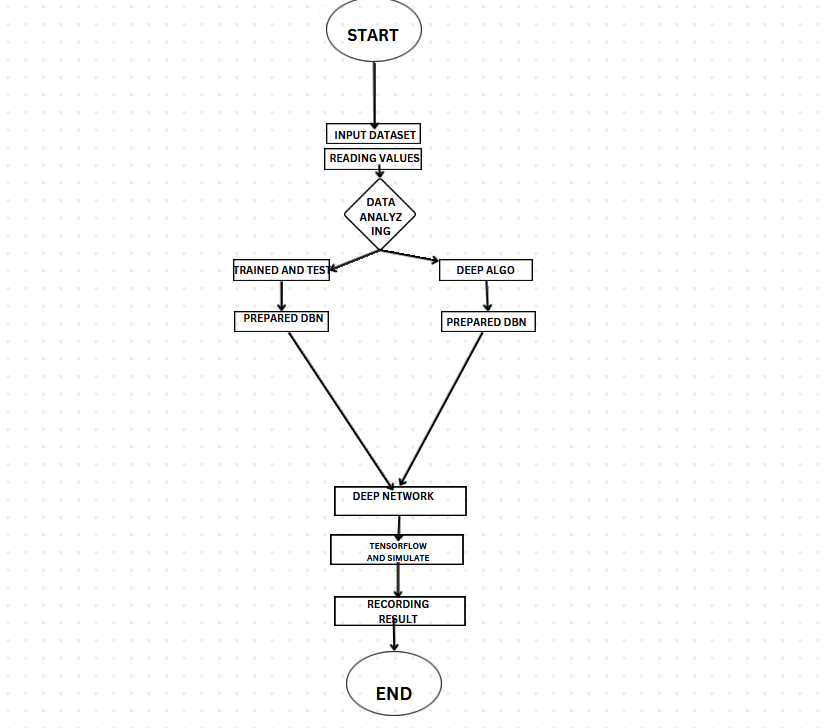
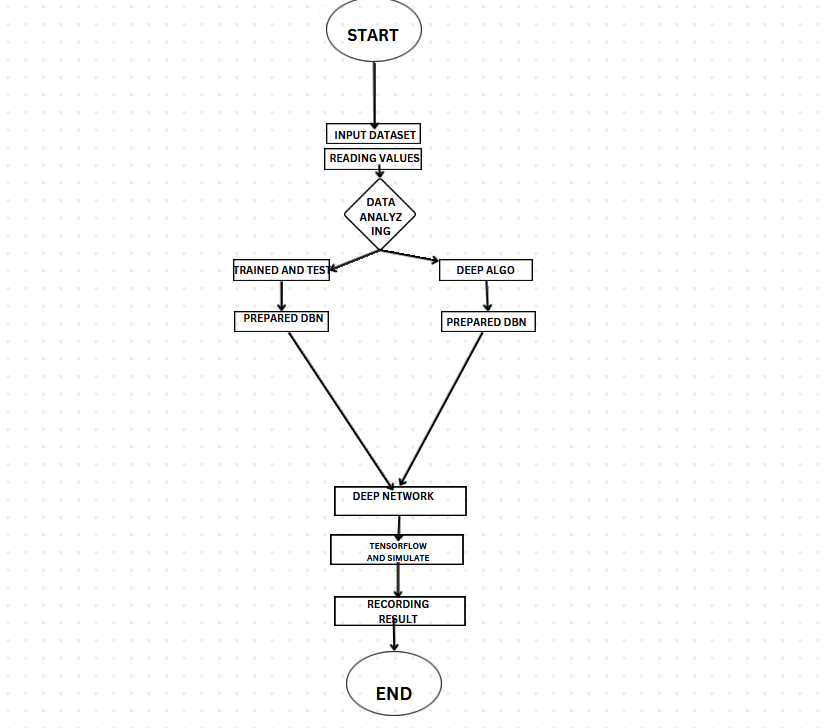
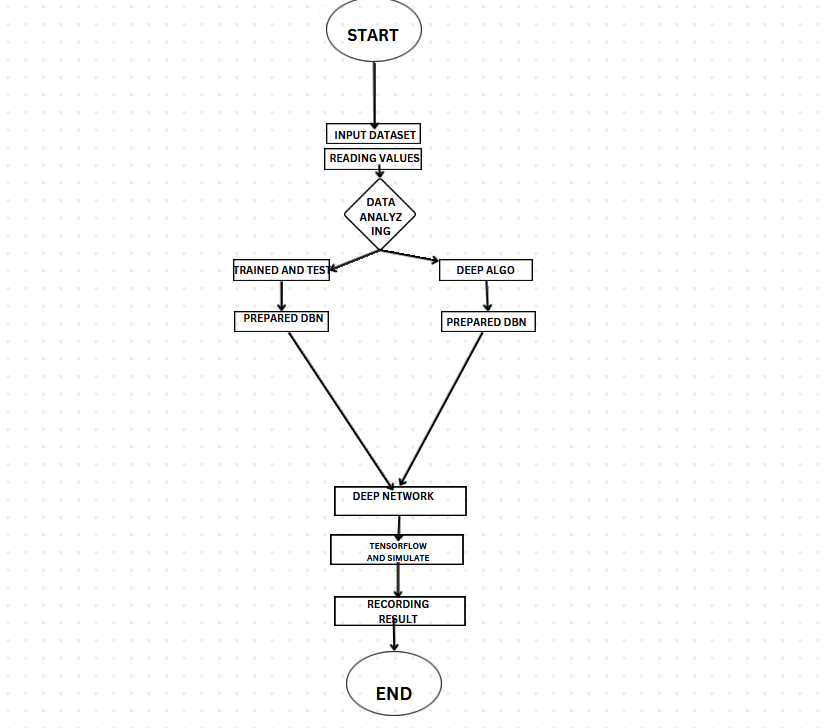
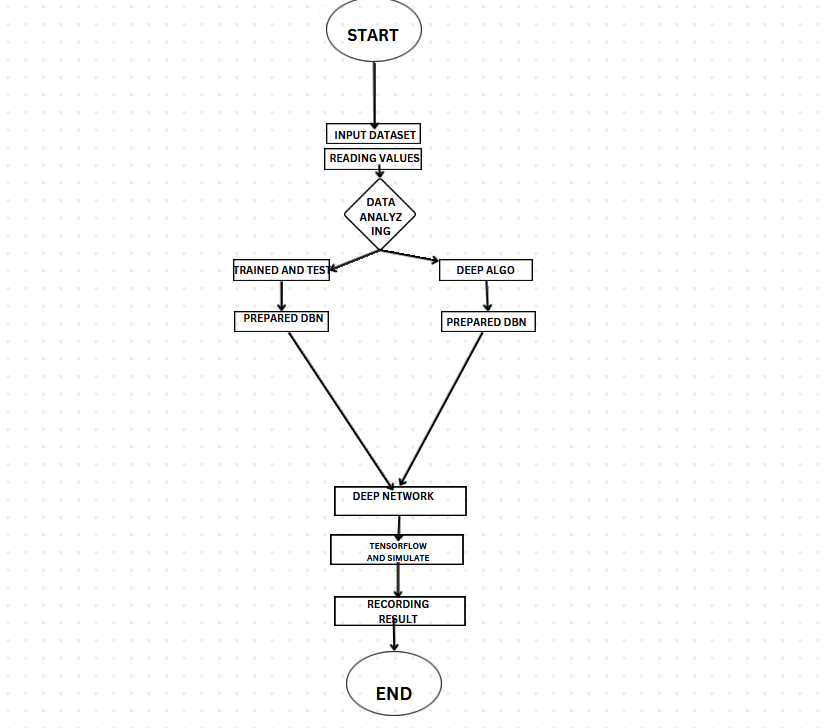
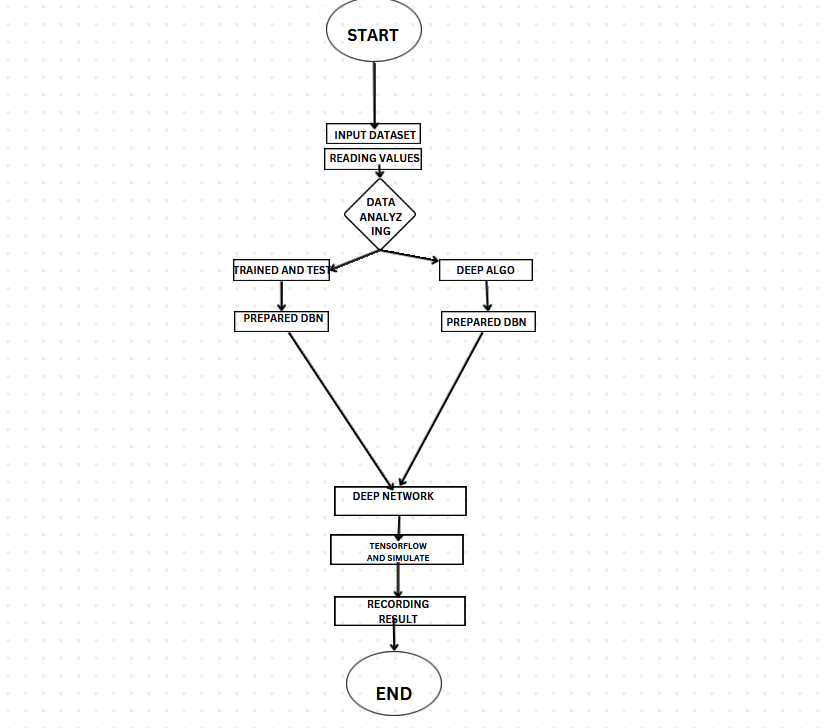
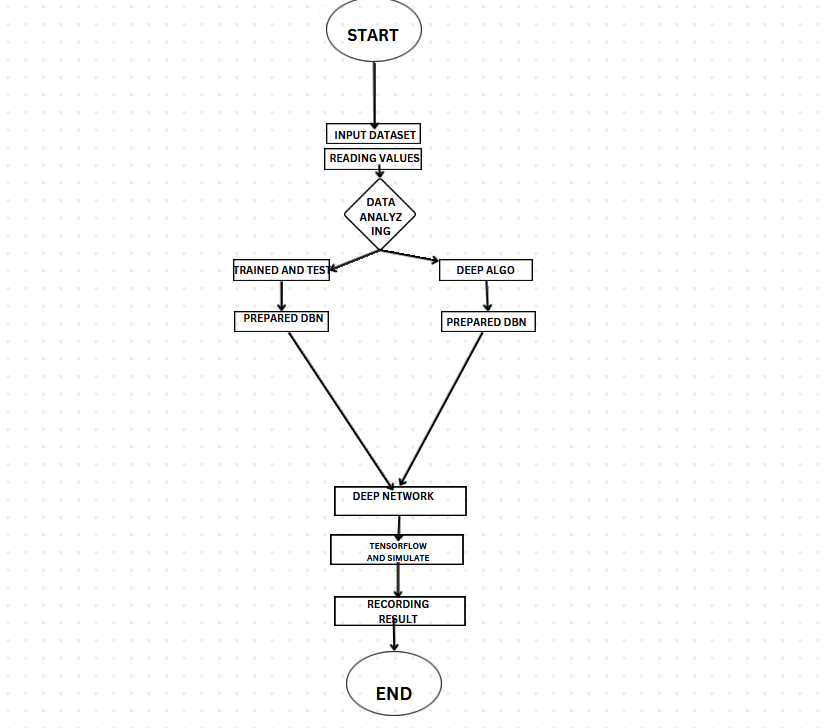
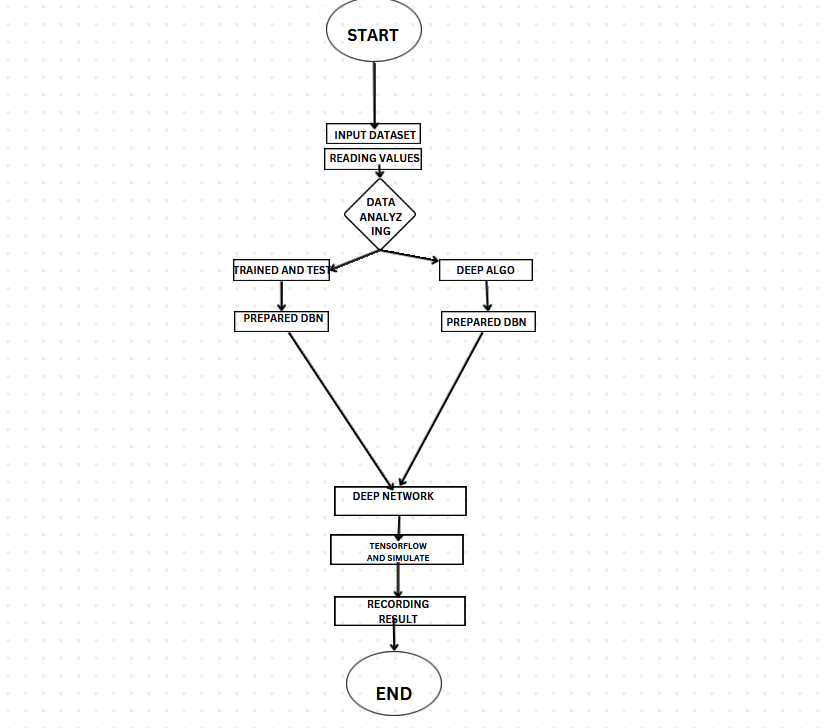
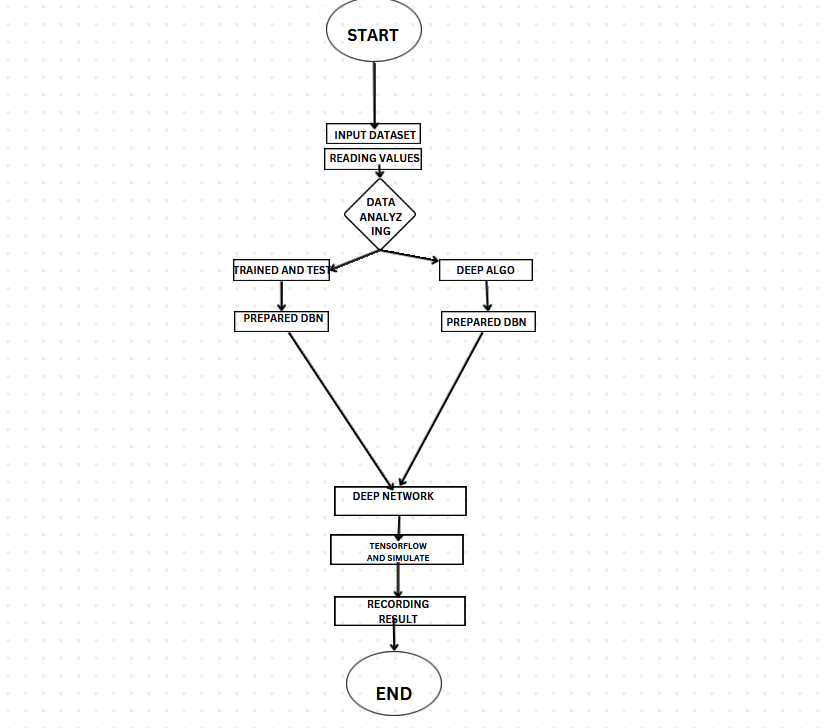
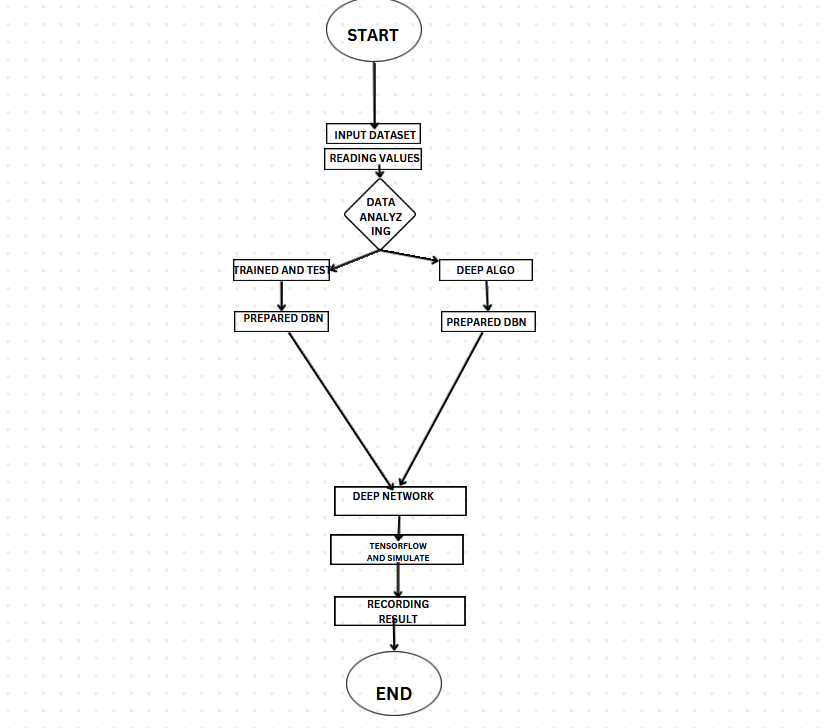
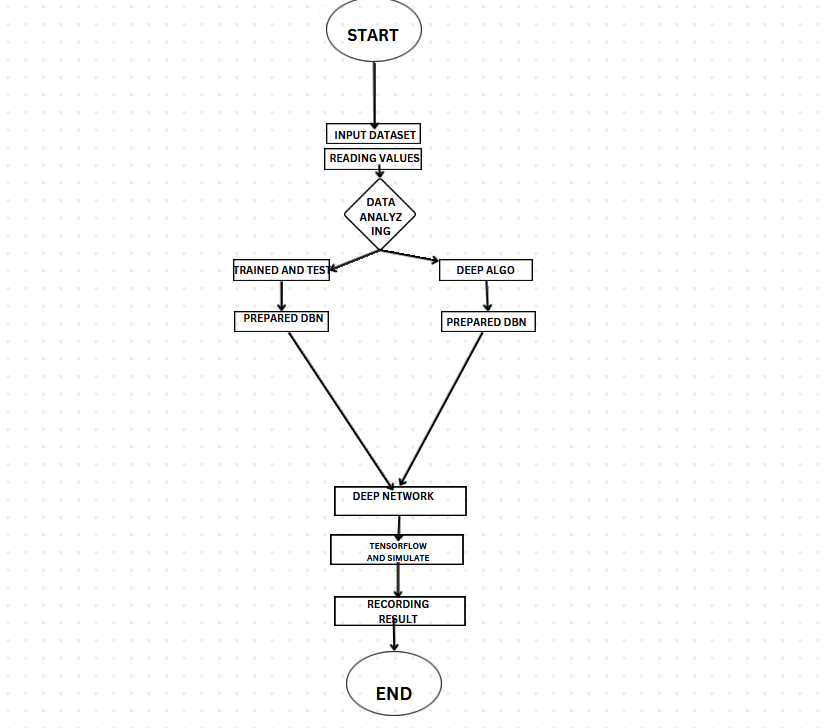
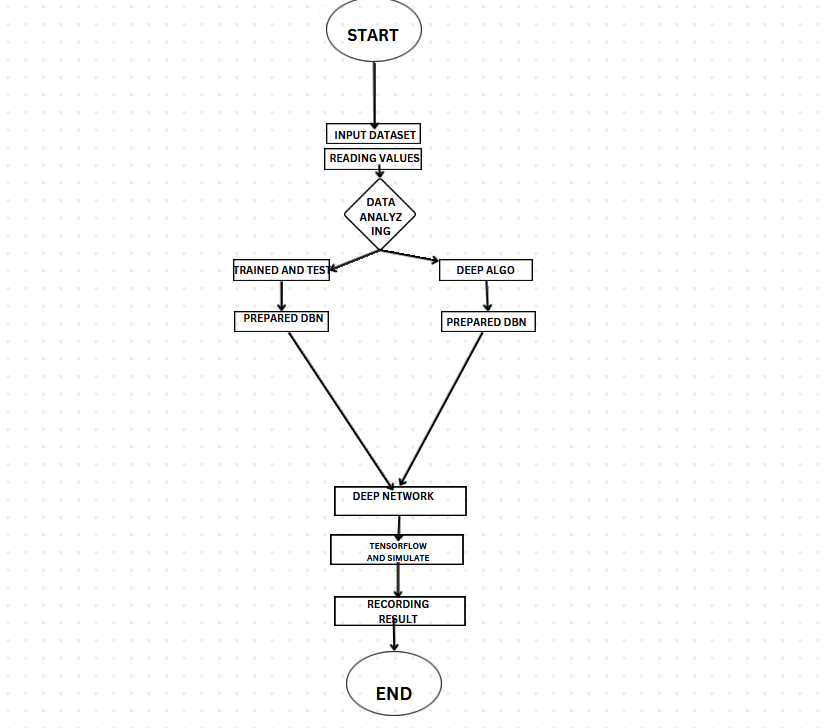
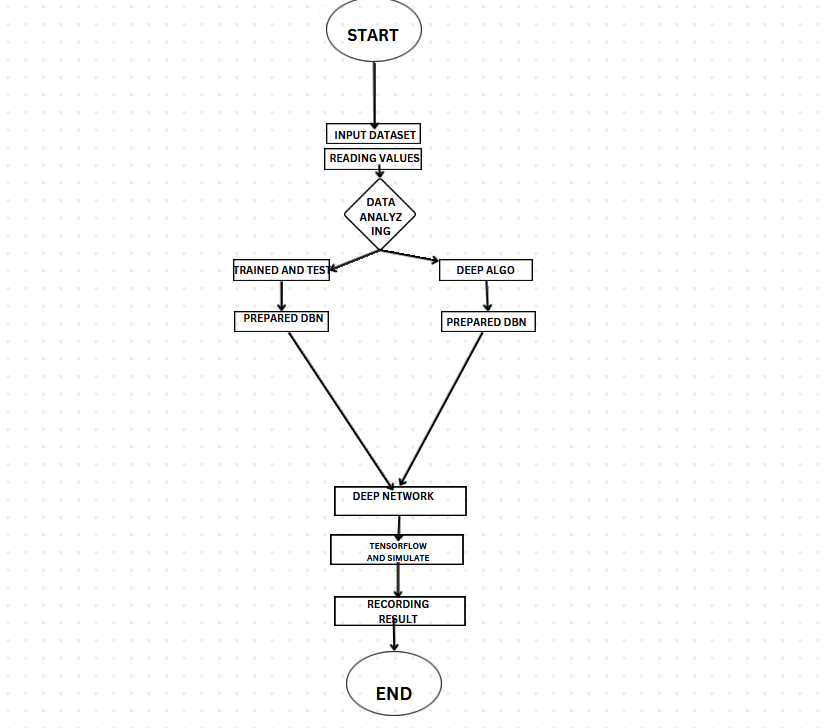
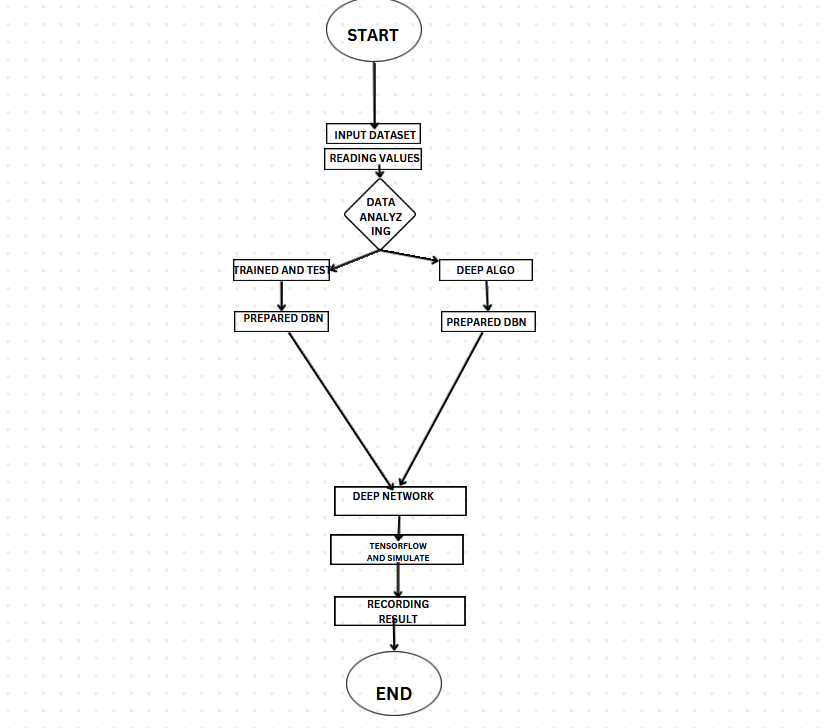
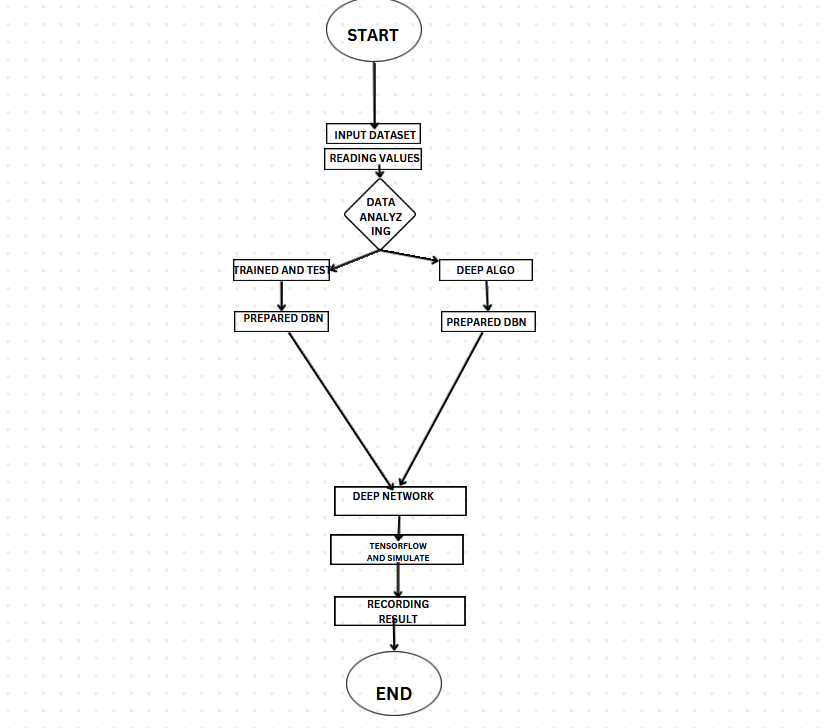
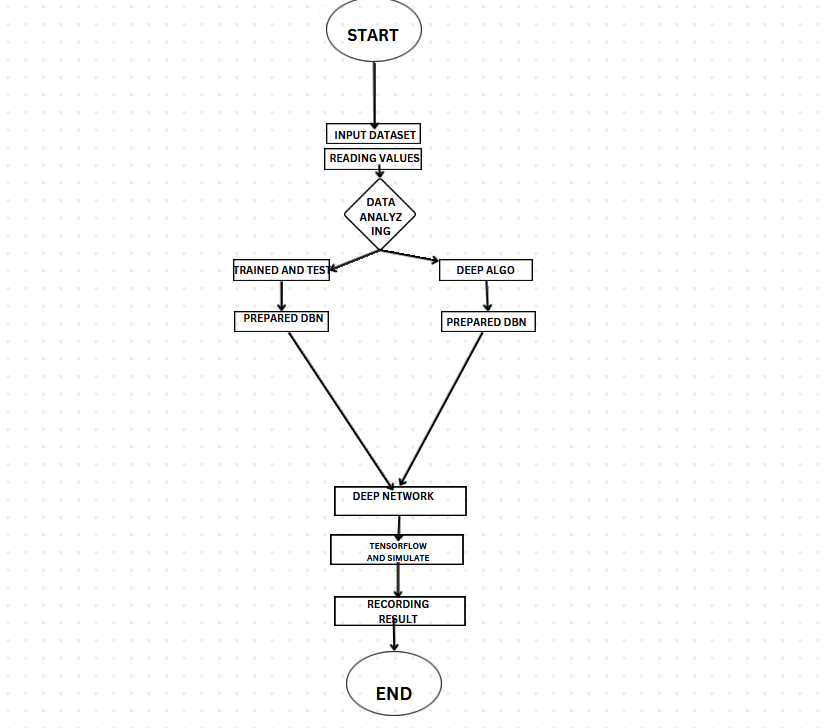
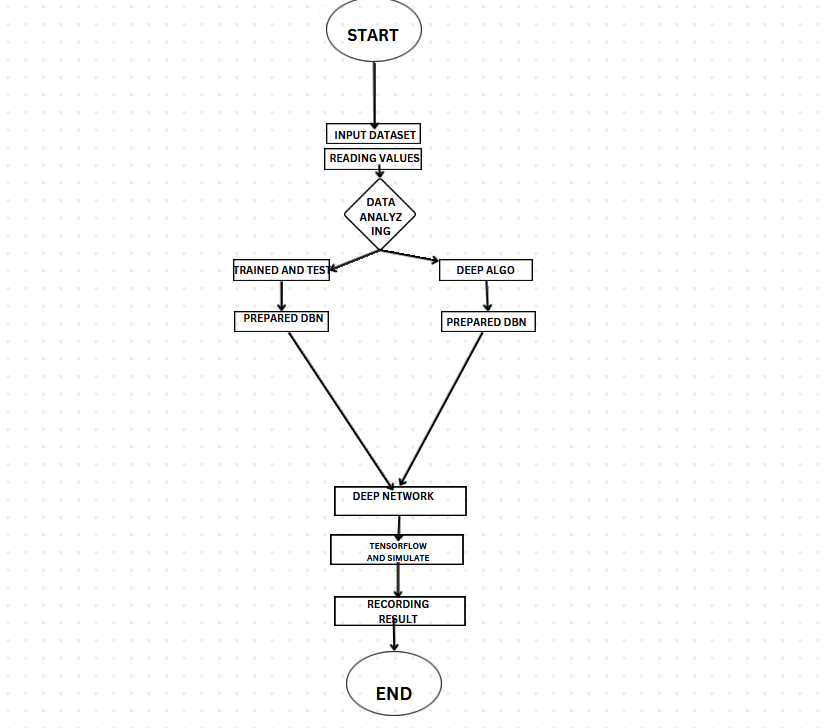
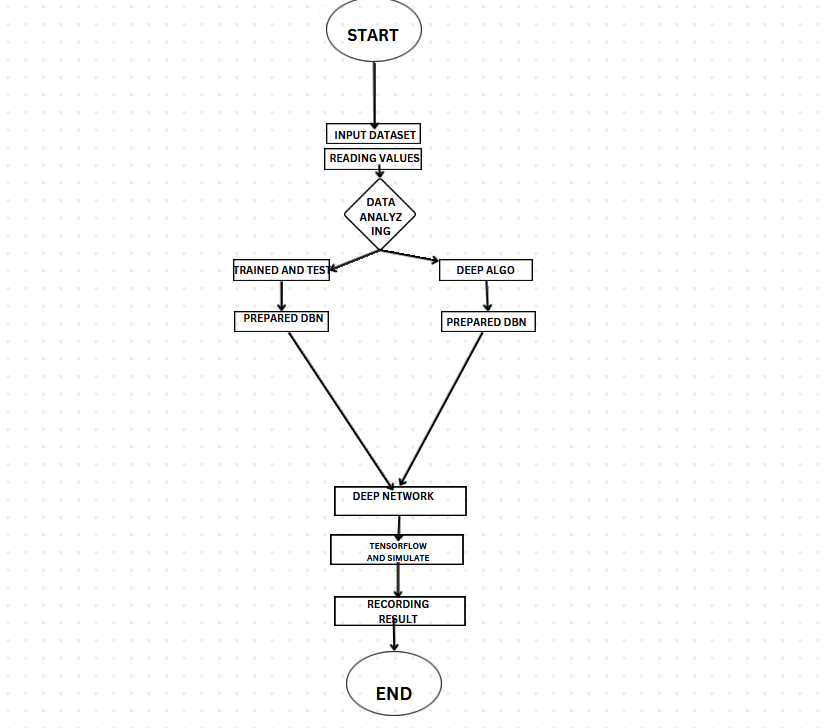
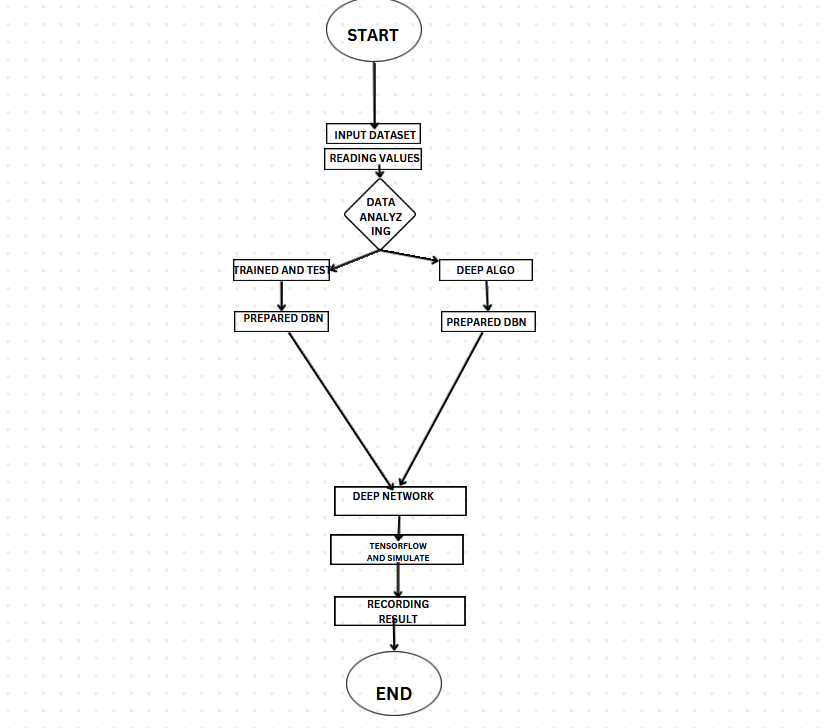
                 

FIG :**THE DBN MODEL ILLUSTRATION**

INPUT layers: Takes the features as input.

HIDDEN layers: Multiple layers of neurons, possibly with dropout for regularization.

OUTPUT layers: Produces a binary prediction using a sigmoid activation function.

TRAINING: Similarly to ANN, focusing on deep architecture and possibly performing the unsupervised training of layers prior of the testing of the dataset which was followed by the supervised fine tuning.

IV. IMPLEMENTATION

This work involves the building and evaluation of three different neural networks algorithm and models for the prediction of cervical cancer in women.

I. Artificial neural Networks (ANN)

II. Long-short Term Memory (LSTM) Network

III. Deep Belief Network (DBN)

There are various technologies and libraries that are used while prediction of the cervical cancer:

The program defined was designed using Python so various libraries that were need to be installed are:

* NumPy, pandas, Seaborn, TensorFlow, Scikit-learn, Imbalance-learn (used for Smoke)
* Tools used: visual studio (VS code) which is capable of installing and running all these libraries.

The dataset that was used in the study is Cervical Cancer Behavior Risk dataset. Which can be found at [*https://www.kaggle.com/datasets/khuzaimaaziz/cervical-cancer-dataset-csv?select=risk\_factors\_cervical\_cancer.csv*](https://www.kaggle.com/datasets/khuzaimaaziz/cervical-cancer-dataset-csv?select=risk_factors_cervical_cancer.csv)

with 35 columns and 836 records of different people of age ranging from 14 to 49 years of age collected from a healthcare or medical research database which contains both numerical and categorical features related to various patients attributes and health indicators ,but there are many missing values .Hence the work developed involves the handling of missing values and converting the categorical data into numerical data and applying the SMOTE for handling the class imbalance in the data which is termed as the pre processing of data in the research domain. The various steps followed while developing the models involves:

Starting with loading of the dataset and then cleaning of data using pandas, the data is first loaded and the missing values were replaced and dropping of the irrelevant columns in the dataset, separating the features labelled as ‘X’ and the target columns labelled as ‘Y’.

**DATA SPLITTING AND STANDARDIZATION**

Splitting of the dataset into the training and testing sets using the **‘train-test split’.** standardization of data is done using **‘StandardScaler’.**

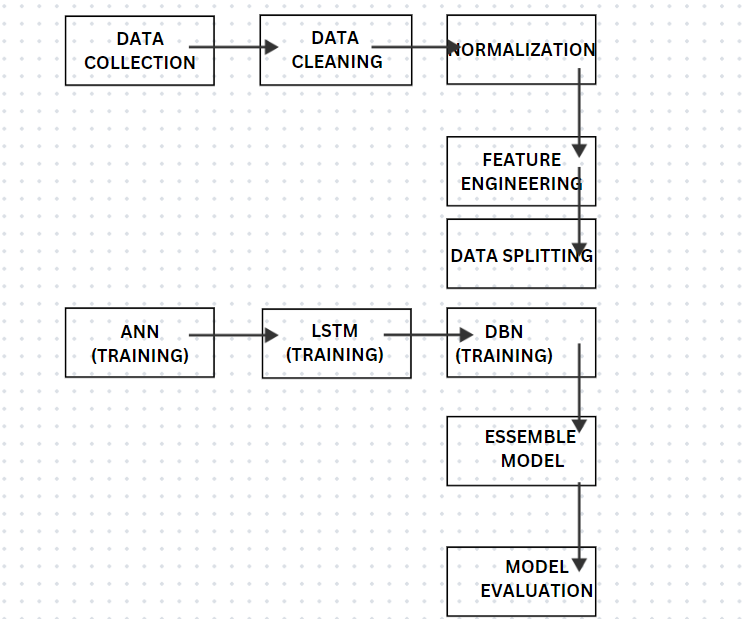
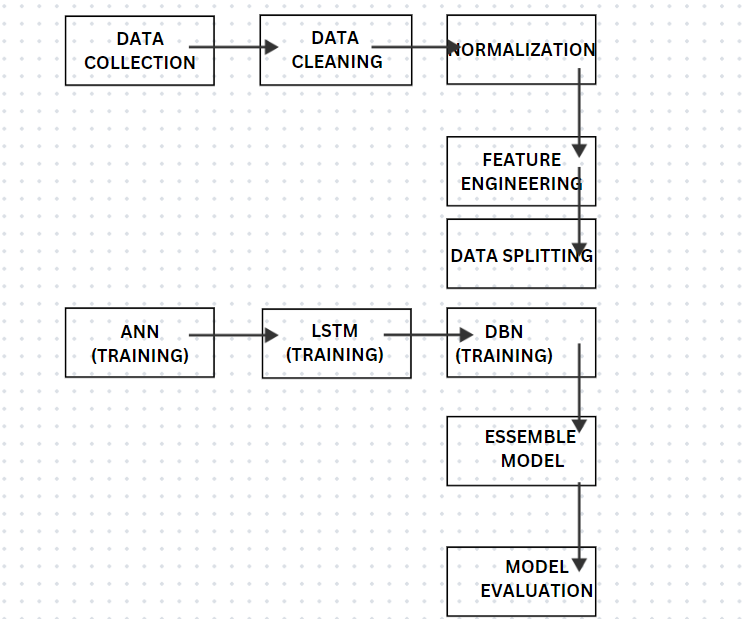
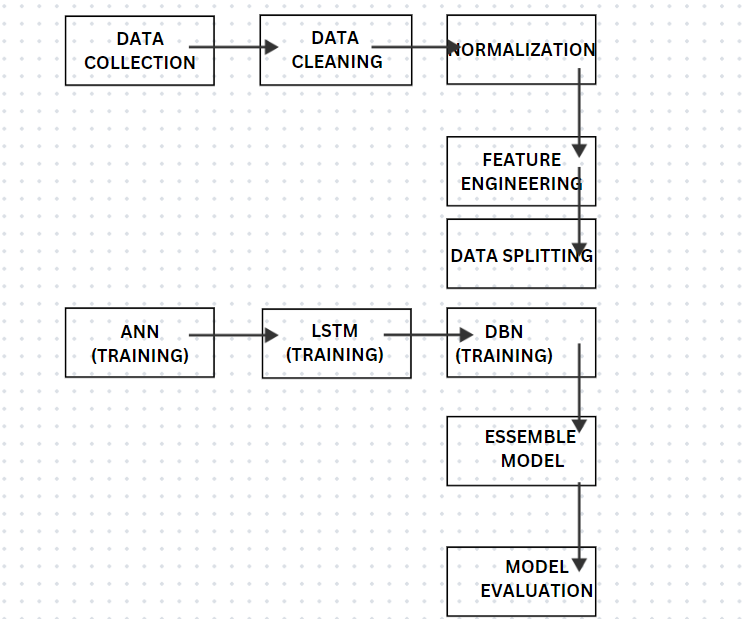
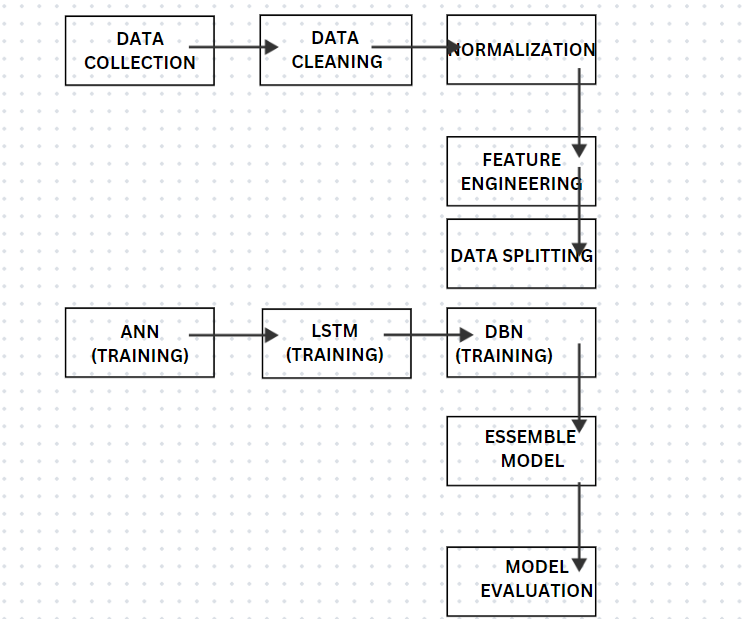
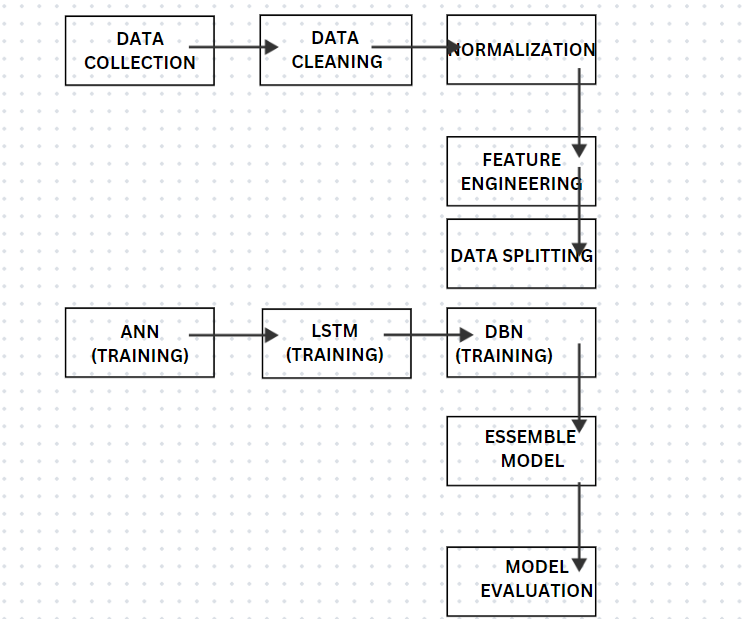
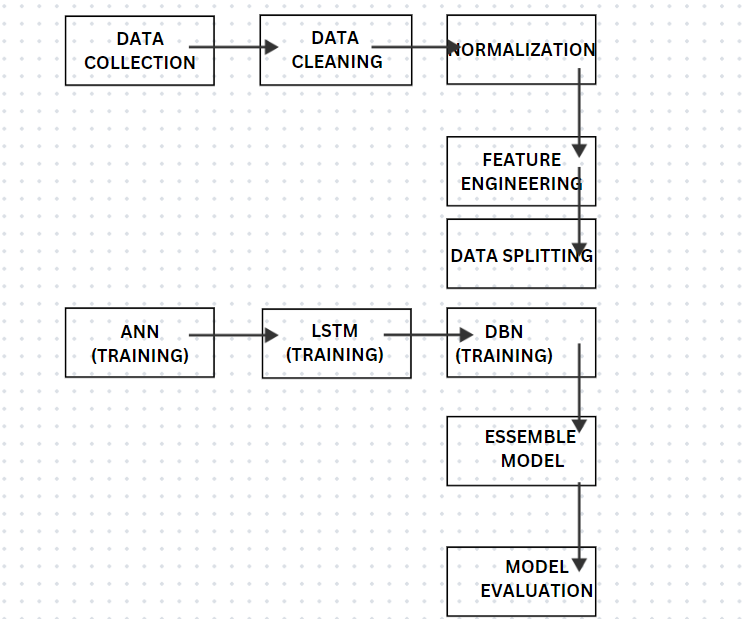
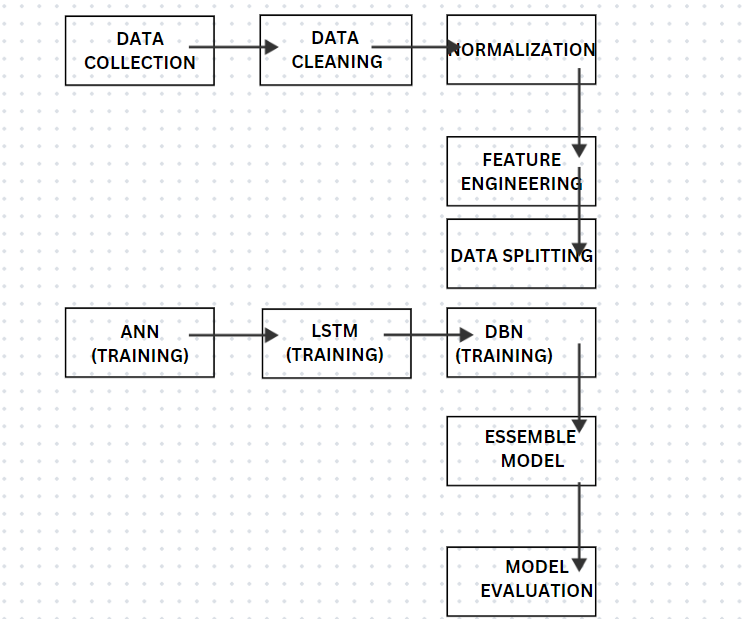
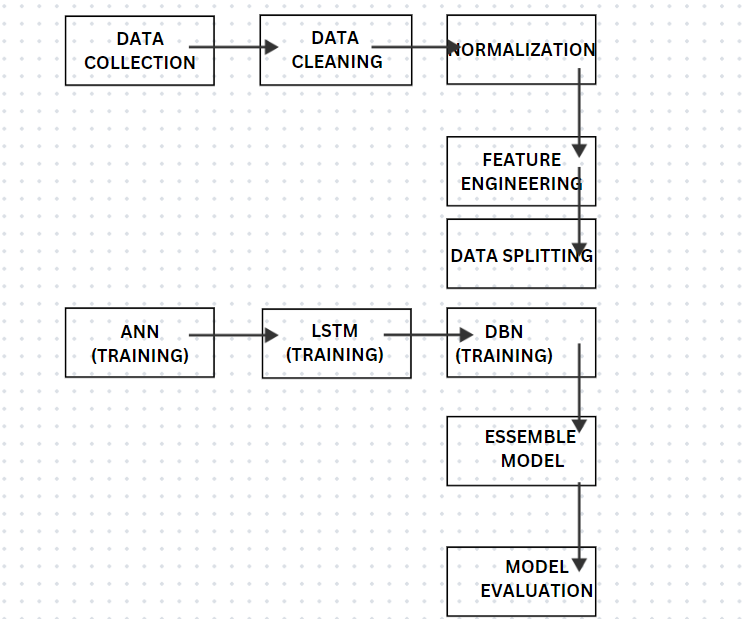
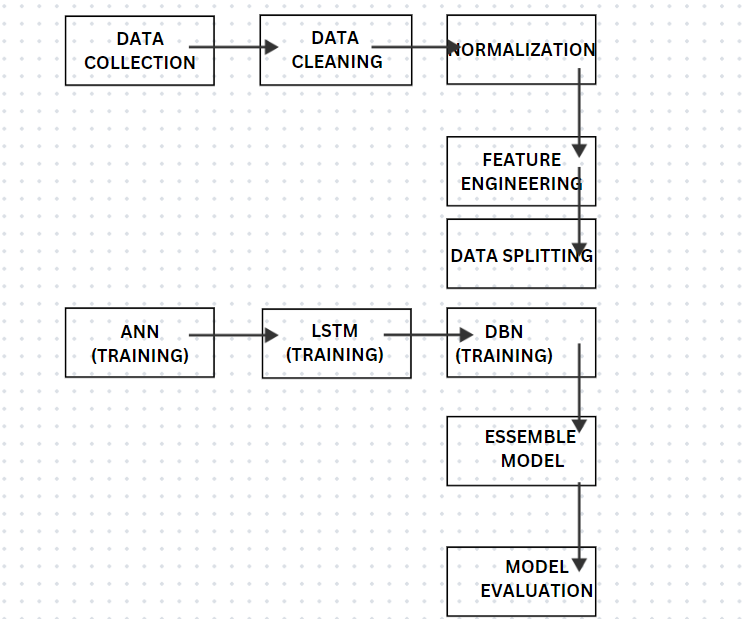
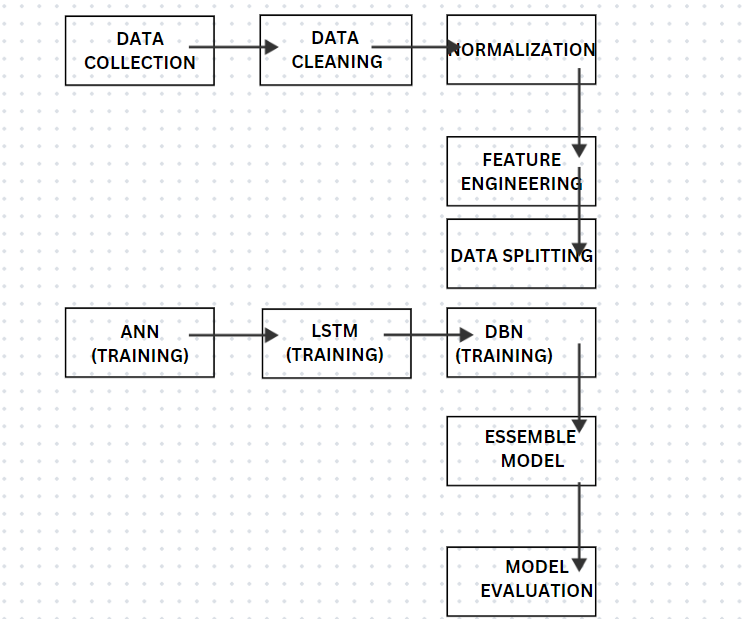
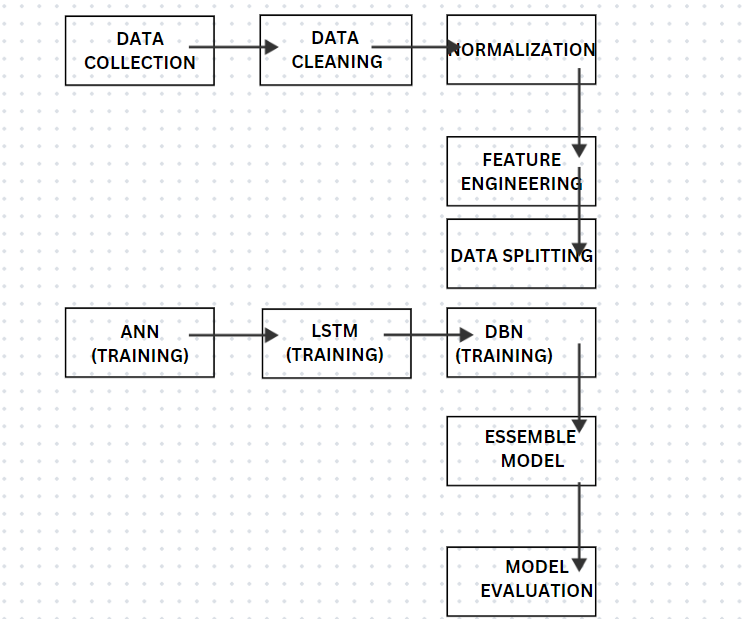
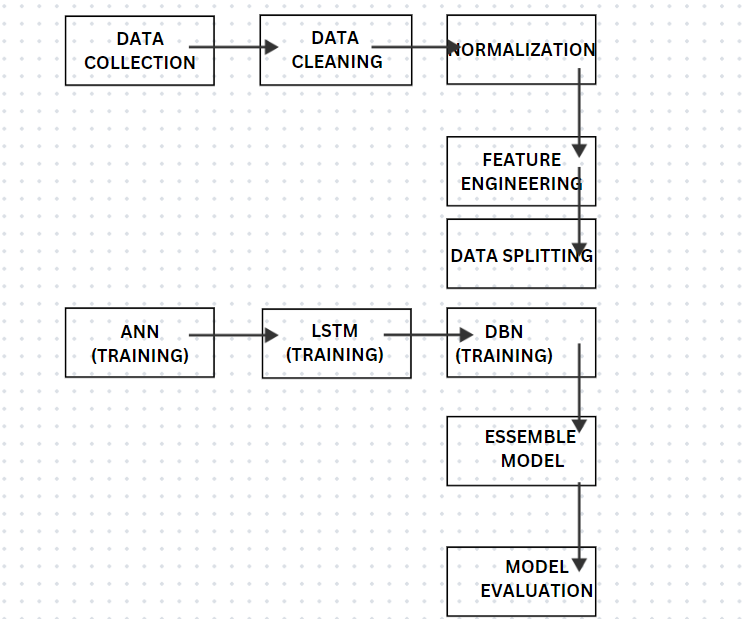
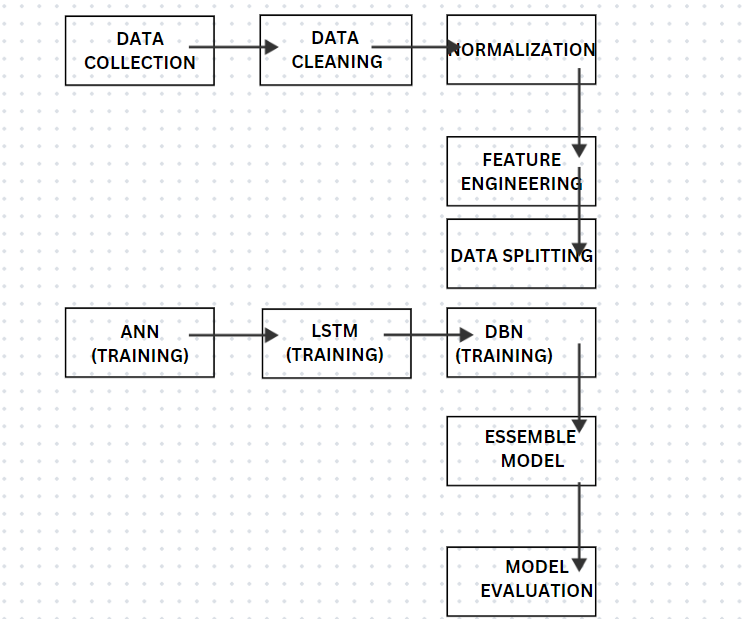
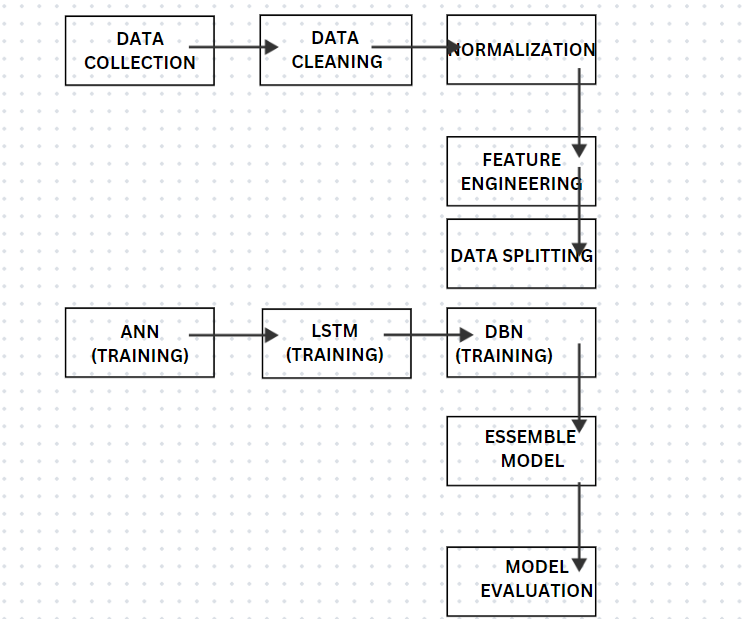
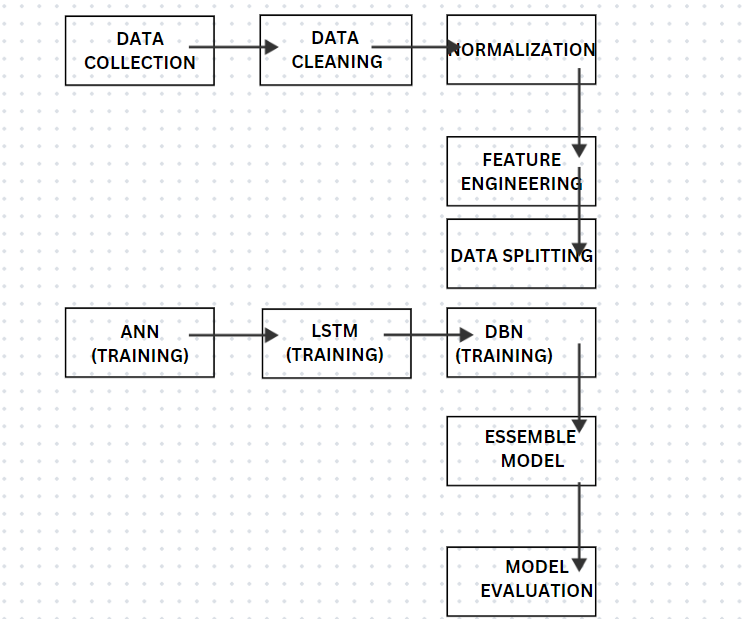
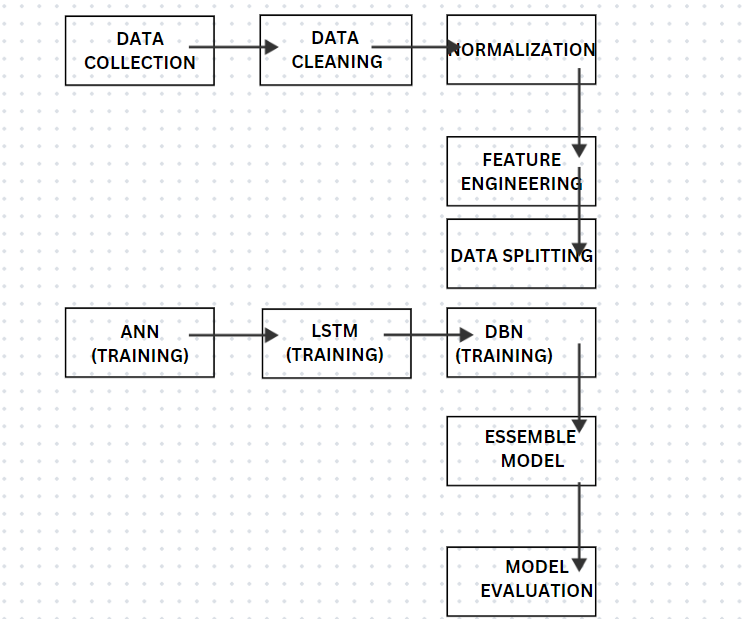
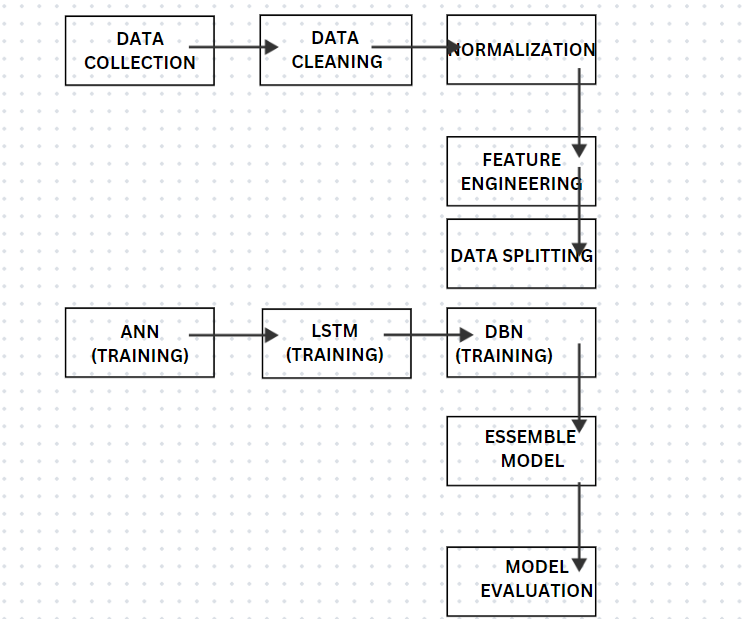
**MODEL TRAINING:**

**I.**ANN: The number of neurons based on the number of the features in the dataset are loaded in the input layer 33 in this case. The hidden layers are loaded with 500 and 200 neurons using ReLU activation. Output layer with 1 neuron and sigmoid activation (binary classification). Compiling and binary cross-entropy loss. The model is trained for **10 epochs and 64 batch size.** Plot the accuracy and loss accuracy score, classification report and confusion matrix using Heatmaps.

**II.** LSTM: the shape of the dataset for time-series by

**(‘(1, x\_train.shape[1])’).** 64 units and dropout of 0.5 output layer with sigmoid activation alike of ANN. compiling of the model is similar to the ANN the data is reshaped on training sets using **(X\_train\_lstm and X\_test\_lstm).** All the evaluation matrix remains the same as in case of ANN.

**III.** DBN: designed using ANN as proxy. hence the architecture defined will be same as that of ANN**.** No specific DBN layers were used to avoid complexity in the model.



**V.RESULTS AND DISCUSSION**

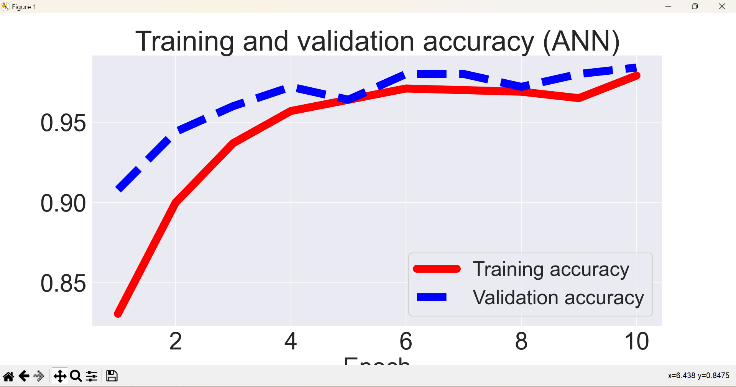
The dataset is in the form tabulated below:

|  |  |
| --- | --- |
| FEATURES | DETAILS |
| Number of instances | 858 |
| Number of attributes | 35 |
| Missing values handling | Filled with mean for numerical, 1.0 for categorical |
| SMOTE | yes |
| Features used | Various numerical and categorical features |
| Target variable | Biopsy (binary classification) |

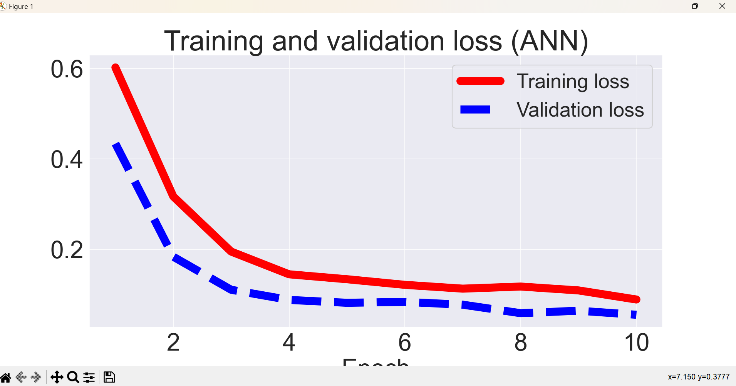
The results of the study is illustrated below in a tabular form:

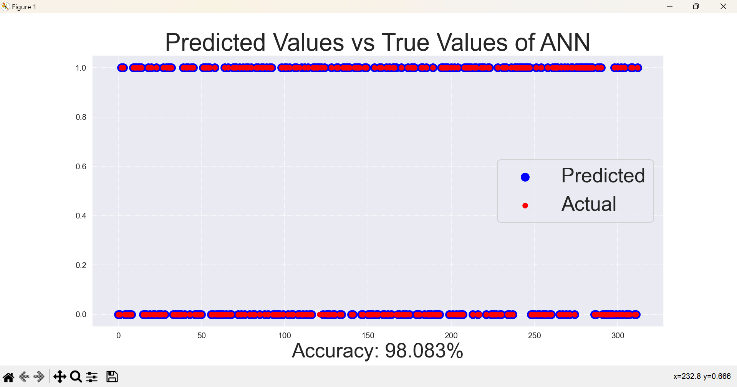
|  |  |  |  |
| --- | --- | --- | --- |
| MODEL NAME | PRECISON | F1 SCORE | ACCURACY |
| ANN | 0.97 | 0.98 | 98.08% |
| LSTM | 0.94 | 0.94 | 94.24% |
| DBN | 0.97 | 0,98 | 98,08% |

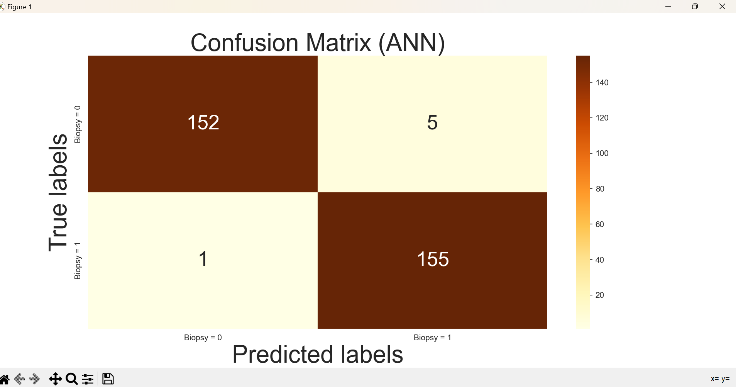
**ANN:**



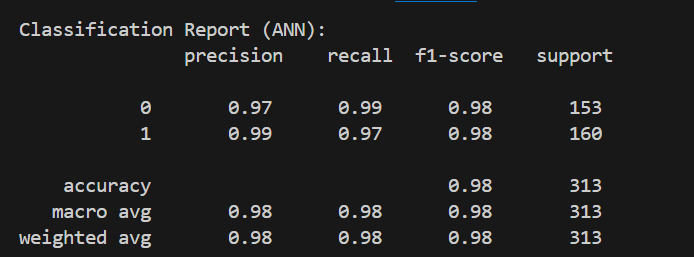
**Fig 1. Trained and validated scores for ANN**

 **Fig 2 The loss employed during the training and testing period**

 **Fig 3 the difference of predicted and true values**

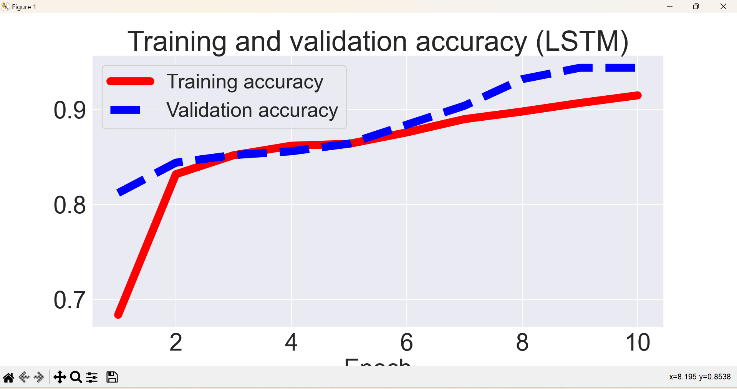
 **Fig 4 the confusion matrix between the true and predicted labels of ANN**

The first graph [Fig 1] explains the graph showing the trained and validated accuracy of the dataset using ANN model. further the graph second [Fig 2] shows the trained and validated loss of the dataset. The illustration of the predicted vs true values of the dataset predicted by the model is shown in fig 3 with an accuracy of 98.083%. Confusion matrix is shown by fig 4

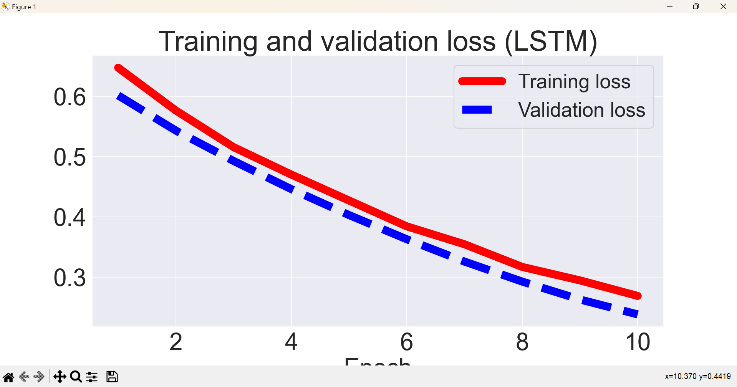


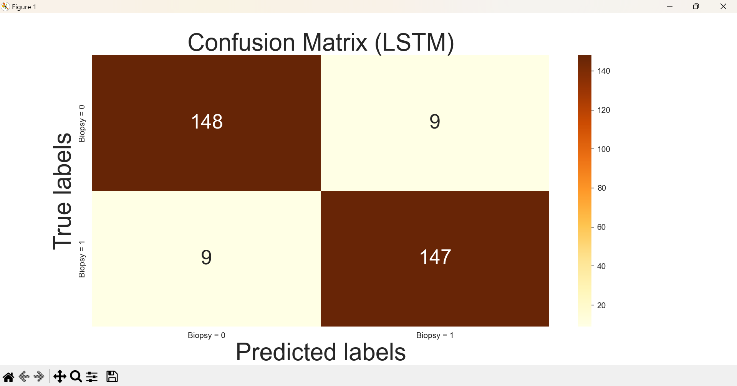
Accuracy of ANN

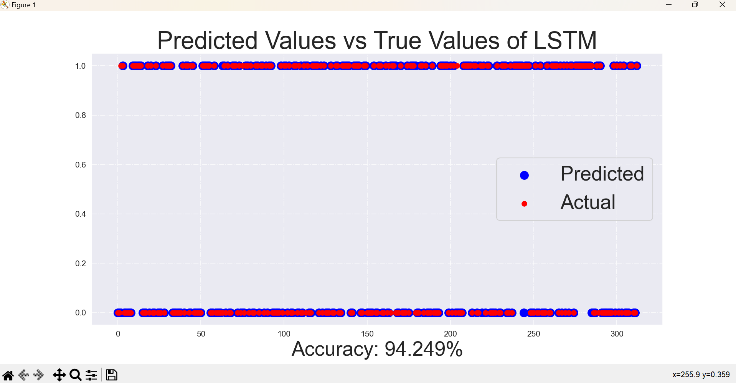
**LSTM:**



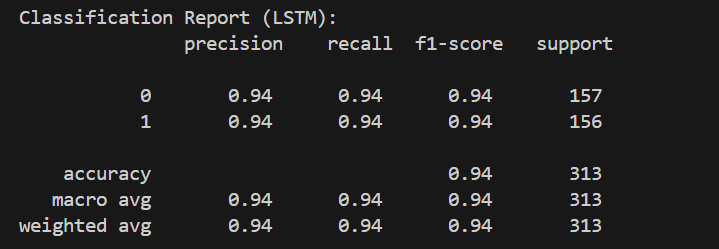
**Fig 5 The trained and validated scores for LSTM**

 **Fig 6 The loss during the trained and validated loss**

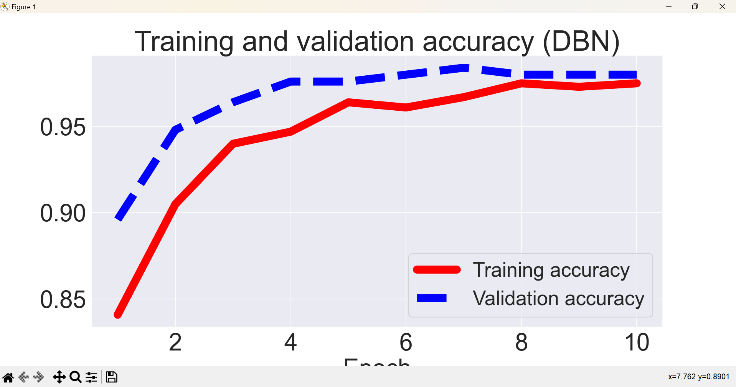
 **Fig 7 The Confusion matrix between the true and the predicted labels**

 **Fig 8 The difference between the predicted and true labels of LSTM**

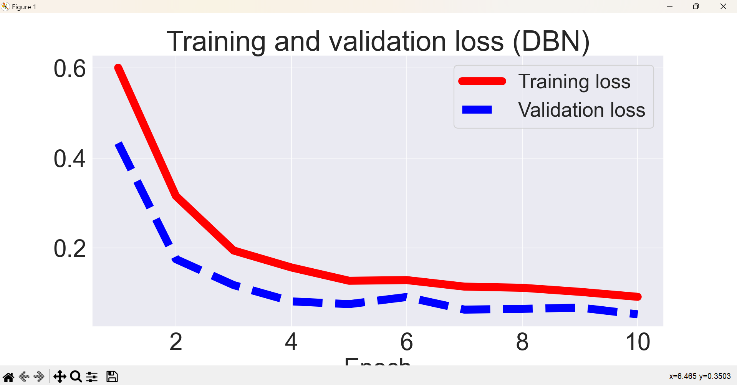
The first graph[Fig 5] for LSTM explains the graph showing the training and validation accuracy of the dataset using LSTM model. further the [fig 6] shows the the loss during the trained and validated phase of the dataset. The illustration of the predicted vs true values of the dataset predicted by the model is shown in [fig 7] with an accuracy of 94.249%. Confusion matrix is shown by [fig 8].

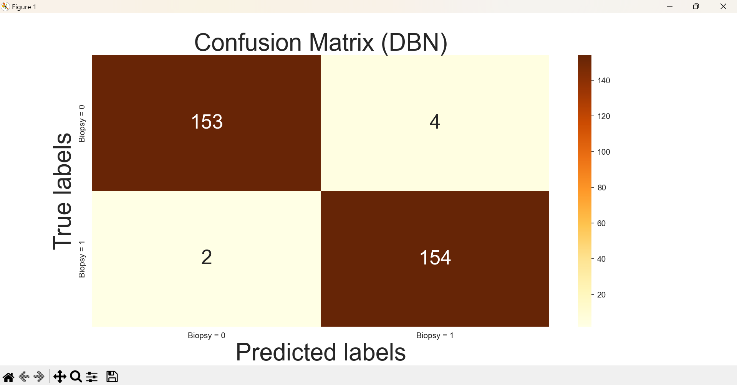


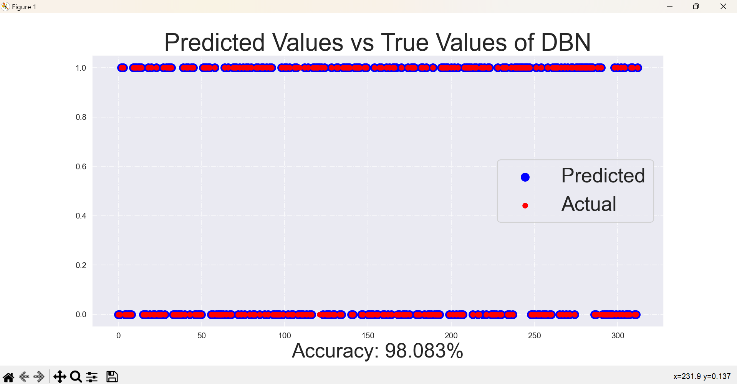
**DBN:**



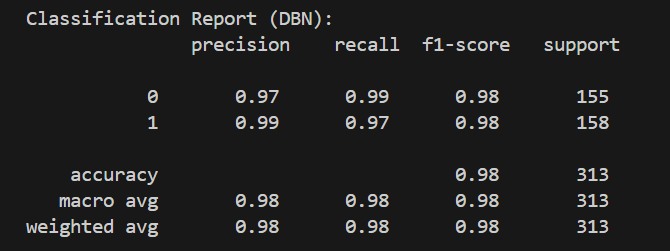
**Fig 9 The trained and validated scores for DBN**

 **Fig 10 The loss during the trained and validated loss**

 **Fig 11 The Confusion matrix between the true and the predicted labels DBN**

 **Fig 12 The difference between the predicted and true labels of DBN**

**The**  first graph for the DBN explains the graph showing the trained and validated accuracy of the dataset using DBN model. further the fig 10 shows the the loss during the trained and validated phase of DBN of the dataset. The illustration of the predicted vs true values of the dataset predicted by the model is shown in fig 11 with an accuracy of 98.083%. Confusion matrix is shown by fig 12



**ANN and DBN** outperformed **LSTM** in terms of accuracy.**ANN** and **DBN** showed similar performance, with DBN leveraging ANN's architecture for a comparable outcome.

**VI. FUTURE ENHANCEMENT**

For further extension of research we will incorporate various other features like:

Feature engineering can be employed to drive more meaningful and useful features from the dataset

Hyperparameter Tuning: This can be used employing grid search or Bayesian optimization to increase the accuracy

Ensemble methods can be used such as boosting, stacking to combine the prediction done by various models for higher accuracy.

**VII.CONCLUSION**

In conclusion the work designed was to predict the cervical cancer in women biopsy using machine learning models like ANN, LSTM, DBN mainly using ANN architecture The comparison between the models were made using the accuracy scores, confusion matrices, and classification reports: ANN is found to be the most robust in case of performance with an accuracy of 98.083% showing the effective use in the binary classification. LSTM is also used in the prediction because of its ability of capturing the dependencies in data. DBN which is build on the ANN architecture performed with higher accuracy which shows the feasibility of using ANN as a building block for more complex algorithm.

Overall, ANN and DBN shows the highest accuracy because of its potential avenues for further exploration and optimization in model architecture and feature engineering. LSTM exhibited the competitive results cause of its suitability for tasks requiring temporal dependencies.

viii.REFERENCES

[1] Hemalatha, B., Karthik, B., Reddy, C. K., & Gokulakrishnan, D. (2024, January). Cervical Cancer Classification: Optimizing Accuracy, Precision, and Recall using SMOTE Preprocessing and t-SNE Feature Extraction. In *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1-7). IEEE.

[2] Tanimu, J. J., Hamada, M., Hassan, M., & Ilu, S. Y. (2021). A contemporary machine learning method for accurate prediction of cervical cancer. In *SHS Web of Conferences* (Vol. 102, p. 04004). EDP Sciences.

[3] Avanija, J., Sunitha, G., Madhavi, K. R., Chakravarthi, S. S., & Vittal, R. H. S. (2022, September). Prediction and analysis of cervical cancer: an ensemble approach. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1734-1739). IEEE.

[4] Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., K S, S., & Sengupta, S. (2022). Predicting cervical cancer biopsy results using demographic and epidemiological parameters: a custom stacked ensemble machine learning approach. *Cogent Engineering*, *9*(1). <https://doi.org/10.1080/23311916.2022.2143040>

[5] Srivastav, S., Guleria, K., & Sharma, S. (2023, March). Predictive Machine Learning Approaches for Cervical Cancer Detection: An Analytical Comparison. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 951-956). IEEE.

[6] Girdonia, M., Garg, R., Jeyabashkharan, P., & Minu, M. S. Cervical Cancer Prediction using Naïve Bayes Classification.

[7] Shpaner, L., Zhang, A., & Singh, K. Predicting Cervical Cancer From Biopsy Results.

[8] Begum, A., & Naveen, S. (2023, November). A Comparative Study on Cervical Cancer Prediction using Various Data Mining Algorithms. In *2023 IEEE North Karnataka Subsection Flagship International Conference (NKCon)* (pp. 1-5). IEEE.

[9] Liu, W., Jin, X., Li, J., Xue, Y., Li, Y., Qian, Z., ... & Yan, X. (2021). Study of cervical precancerous lesions detection by spectroscopy and support vector machine. *Minimally Invasive Therapy & Allied Technologies*, *30*(4), 208-214.

[10] Ali, M. M., Ahmed, K., Bui, F. M., Paul, B. K., Ibrahim, S. M., Quinn, J. M., & Moni, M. A. (2021). Machine learning-based statistical analysis for early stage detection of cervical cancer. *Computers in biology and medicine*, *139*, 104985.

[11] Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., ... & Hamdi, M. (2022). Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. *Computational and Mathematical Methods in Medicine*, *2022*(1), 4688327.