NAME: Tarani Chilamkoti

NJIT UCID: tc533

Email Address: tc533@njit.edu

Date: 20th March 2025

Professor: Yasser Abduallah

SP25-CS634854 Data Mining

Final Project Report

**Breast Cancer Classification using LSTM, Random Forest, and Decision Tree**

**Abstract:**

This project explores binary classification on the Breast Cancer Wisconsin dataset using three algorithms: Random Forest, LSTM (Long Short-Term Memory network), and Decision Tree. The primary goal is to compare these models' performance in predicting whether a tumor is malignant or benign. The project measures accuracy, precision, recall, F1-score, and other metrics using 10-fold cross-validation.

**Introduction:**

Breast cancer diagnosis is a critical medical challenge. Data mining and machine learning techniques provide automated, reliable classification methods that can assist doctors in decision-making. This project compares traditional classifiers (Random Forest, Decision Tree) with a deep learning approach (LSTM) to evaluate their effectiveness on the Breast Cancer Wisconsin dataset.

**Core Concepts and Principles:**

* **Random Forest:** An ensemble learning method that constructs multiple decision trees and outputs the majority vote for classification.
* **LSTM:** A type of recurrent neural network (RNN) designed to handle sequential data and retain long-term dependencies — useful for recognizing patterns in medical data.
* **Decision Tree:** A straightforward, interpretable model that splits data into branches based on feature conditions.

**Project Workflow:**

1. **Data Loading and Preprocessing:**
   * Dataset is loaded from UCI’s repository.
   * The 'ID' column is dropped as it’s non-informative.
   * Labels ('M' for malignant, 'B' for benign) are encoded to binary values (1 and 0).
   * Features are standardized using **StandardScaler** for better model performance.
2. **Model Training:**
   * Each algorithm undergoes **10-fold cross-validation** to ensure generalizability.
3. **Performance Metrics:**
   * Accuracy
   * Precision
   * Recall
   * F1 Score
   * False Positive Rate (FPR)
   * False Negative Rate (FNR)
   * True Skill Statistic (TSS)
   * Heidke Skill Score (HSS)

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**Algorithm Implementation:**

1. **Random Forest:**
   * 100 estimators (trees) are trained.
   * Random subsets of data are sampled for training.
2. **LSTM:**
   * Sequential model with one LSTM layer (64 units) and dropout (20%) to prevent overfitting.
   * The output layer uses a sigmoid activation function for binary classification.
   * Compiled with Adam optimizer and binary cross-entropy loss.
   * Trained for 10 epochs with a batch size of 8.
3. **Decision Tree:**
   * Gini impurity is used to measure splits.
   * Unrestricted depth for maximum performance.

**Results and Evaluation:**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

The performance metrics were averaged across 10 folds. Below is the summary of results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Random Forest** | **LSTM** | **Decision Tree** |
| Accuracy | 96.5% | 94.8% | 92.3% |
| Precision | 97.2% | 93.5% | 91.8% |
| Recall | 95.8% | 94.3% | 92.1% |
| F1 Score | 96.5% | 93.9% | 92.0% |
| FPR | 3.4% | 5.2% | 7.7% |
| FNR | 4.2% | 5.7% | 7.9% |
| TSS | 0.926 | 0.889 | 0.842 |
| HSS | 0.918 | 0.870 | 0.821 |

**Conclusion:**

This project demonstrates that ensemble methods (Random Forest) outperform both deep learning (LSTM) and traditional (Decision Tree) approaches for this dataset. LSTM remains promising for larger, more complex datasets, while Decision Trees provide interpretability at the cost of slightly lower performance.

**References:**

* Breast Cancer Wisconsin Dataset: UCI Machine Learning Repository
* Scikit-learn Documentation: https://scikit-learn.org/
* TensorFlow LSTM Tutorial: https://www.tensorflow.org/text/tutorials/text\_classification\_rnn

**Appendix:**

* Source code (.py file) and dataset (.csv file) are included in the final submission.
* Google colab
* Github repo: <https://github.com/TaraniChilamkoti/Data-Mining-FinalTerm-Project.git>

**End of Report**