

A
Project Report on
System Threat Forecaster

Submitted to the Department of Electronics Engineering in Partial Fulfilment of
the Requirement for the AICTE QIP PG Certification Programme on

Deep Learning: Fundamentals and Applications

by

Mr. Milav Jayeshkumar Dabgar

Guided by

Dr. Jignesh N. Sarvaiya, Dr. Kishor Upla,

Dr. Kamal Captain, Dr. Suman Deb

Coordinators- AICTE QIP PG Certification Programme, DECE



DEPARTMENT OF ELECTRONICS ENGINEERING
SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY
DECEMBER-2025

Declaration

- I hereby declare that the work being presented in this Report entitled "**System Threat Forecaster**" by **Milav Jayeshkumar Dabgar**, submitted in partial fulfillment of the requirements for the successful completion of the AICTE QIP PG Certification Programme on "Deep Learning: Fundamentals and Applications" conducted at Sardar Vallabhbhai National Institute of Technology, Surat, during the academic year 2025 – 2026.
- Neither the source code therein nor the content of the report has been copied or downloaded from any other source. I understand that my result grades could be revoked if it is found that they are incorrect later.

Milav Jayeshkumar Dabgar

(Signature of the Candidate)

Date: _____

Sardar Vallabhbhai National Institute Of Technology

Surat - 395 007, Gujarat, India

DEPARTMENT OF ELECTRONICS ENGINEERING



CERTIFICATE

This is to certify that the project entitled '**System Threat Forecaster**' has been successfully completed by **Mr. Milav Jayeshkumar Dabgar**. This report is being submitted in partial fulfillment of the requirements for the successful completion of the AICTE QIP PG Certification Programme on **Deep Learning: Fundamentals and Applications** in Department of Electronics Engineering during the academic year **2025-26**.

Examiner-1

Examiner-2

Examiner-3

Examiner-4

Seal of The Department

DECEMBER-2025

Acknowledgements

We wish to express our deepest sense of gratitude and sincere thanks to all those who have contributed to the successful completion of this project report titled, **System Threat Forecaster** which was undertaken as a requirement for the AICTE QIP PG Certification Programme on "Deep Learning: Fundamentals and Applications."

We are profoundly indebted to the faculty members, especially our guide and mentor for this project, at the Department of Electronics Engineering at Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, Gujarat, for providing the excellent learning environment, technical facilities, and insightful direction that were crucial for this work.

We extend our sincere thanks to the leadership and organizing committee for the successful execution of the QIP course: **Prof. Anupam Shukla, Director, SVNIT Surat (Chair Patron), Dr. Shilpi Gupta, HOD, DOECE, SVNIT Surat (Patron), Prof. A. A. Shaikh, DoME, SVNIT, Surat (Centre Coordinator), Dr. Jignesh N. Sarvaiya, Dr. Kishor Upla, Dr. Kamal Captain, and Dr. Suman Deb (Coordinators from DoECE).**

We also acknowledge the support and sponsorship provided by the All India Council for Technical Education (AICTE) under the Quality Improvement Programme (QIP), which has enabled us to enhance our knowledge and skills in deep learning through this opportunity.

Finally, we express our gratitude to our respective parent institutions, **Government Polytechnic Palanpur**, for granting the necessary support and time to participate in this Program.

Mr. Milav Jayeshkumar Dabgar
Lecturer, Dept of ECE
GP Palanpur
DECEMBER 2025

Abstract

Cybersecurity threats continue to evolve rapidly, with malware infections causing significant damages to organizations worldwide. This project presents a comprehensive comparative study of machine learning and deep learning approaches for malware threat prediction, implemented as the System Threat Forecaster.

The research evaluates thirteen models: seven ML algorithms (Decision Tree, Random Forest, LightGBM, Naive Bayes, Logistic Regression, AdaBoost, SGD) and six DL architectures (Simple MLP, Deep MLP, Residual Network, Attention Network, Wide & Deep, FT-Transformer). Using a dataset of 100,000 samples with 75 features (47 numerical, 28 categorical), we implemented a complete pipeline encompassing pre-processing, feature engineering, model training, and evaluation.

LightGBM achieved the best performance at 62.94% validation accuracy (F1-score: 0.6286), outperforming all deep learning models. The best DL model (Deep MLP) reached 61.79% accuracy with 63,714 parameters. This 1.15% performance gap demonstrates that tree-based ensemble methods maintain superiority over neural networks for tabular data. All DL architectures clustered around 61.5%, confirming that model complexity does not overcome dataset limitations.

Analysis revealed weak feature correlations (max 0.118), establishing a performance ceiling around 63%. The top Kaggle score of 69.6% represents only 6.7% improvement over our baseline. This work includes a production deployment at <https://stf.milav.in> and provides evidence-based guidance for choosing ML over DL for tabular cybersecurity data.

Table of Contents

	Page
Acknowledgements	vii
Abstract	ix
Table of Contents	xi
List of Figures	xiii
List of Tables	xv
List of Abbreviations	xvii
Chapters	
1 Introduction	1
1.1 Background and Motivation	1
1.1.1 The ML vs. DL Debate for Tabular Data	1
1.2 Problem Statement and Research Questions	2
1.2.1 Technical Challenges	2
1.2.2 Research Questions	2
1.3 Objectives and Contributions	3
1.3.1 Primary Objectives	3
1.3.2 Key Contributions	3
1.4 Thesis Organization	4
2 Dataset, Metadata, and Exploratory Data Analysis	5
2.1 Dataset Overview	5
2.2 Dataset Metadata	5
2.2.1 Feature Categories	6
2.3 Exploratory Data Analysis	6
2.3.1 Key EDA Findings	6
2.3.2 Missing Values and Distributions	6
2.4 Modeling Implications	6
3 Experimental Setup and Model Selection	9
3.1 Data Preprocessing	9
3.2 Machine Learning Models	9
3.3 Deep Learning Models	9
3.3.1 Training Configuration	10
4 Experimental Results and Discussion	11
4.1 Machine Learning Results	11
4.1.1 Key ML Findings	12
4.2 Deep Learning Results	12
4.2.1 Key DL Findings	12
4.3 ML vs. DL Comparison	13

Table of Contents

4.3.1	Why ML Outperforms	13
4.4	Kaggle Comparison	14
4.5	Summary	14
5	Conclusion and Future Work	15
5.1	Key Findings	15
5.2	Contributions	15
5.3	Practical Recommendations	15
5.4	Limitations	15
5.5	Future Work	15
5.6	Concluding Remarks	16
References	19

List of Figures

2.1	Target Class Distribution: Near-perfect 50.52% malware vs. 49.48% clean split eliminates need for resampling strategies.	5
2.2	Feature Correlation Heatmap: Maximum correlation with target (Has-Detections) is only 0.118, demonstrating weak linear relationships and necessitating ensemble methods.	7
2.3	Missing Values Analysis: Up to 30% missing data in security software features, requiring careful imputation strategies.	7
4.1	Comprehensive Model Comparison: LightGBM achieves highest validation accuracy (62.94%), with all DL models clustering tightly between 61.45-61.79%.	11
4.2	LightGBM Confusion Matrix: Balanced performance across both classes with 62.94% overall accuracy.	12
5.1	Deployed Web Application: Homepage showcasing System Threat Forecaster with 7 ML and 6 DL models for real-time malware prediction. . .	17
5.2	Model Performance Dashboard: Interactive comparison of all 13 models with detailed metrics and hyperparameters.	17
5.3	Prediction Interface: User-friendly form accepting 75 features for real-time threat prediction using all trained models.	18

List of Tables

2.1	Dataset Metadata	5
4.1	ML Models: Performance Metrics	11
4.2	DL Models: Performance Metrics	13
4.3	Best ML vs. Best DL	13

List of Abbreviations

SVNIT	Sardar Vallabhbhai National Institute of Technology
AICTE	All India Council for Technical Education
QIP	Quality Improvement Programme
ML	Machine Learning
DL	Deep Learning
NN	Neural Network
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
AI	Artificial Intelligence
EDA	Exploratory Data Analysis
API	Application Programming Interface
GPU	Graphics Processing Unit
MPS	Metal Performance Shaders
SGD	Stochastic Gradient Descent
AdamW	Adam with Weight Decay
ReLU	Rectified Linear Unit
IoT	Internet of Things
OS	Operating System
RAM	Random Access Memory
URL	Uniform Resource Locator
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
CSV	Comma-Separated Values

Chapter 1

Introduction

1.1 Background and Motivation

The cybersecurity landscape has undergone dramatic transformation in recent years. Modern organizations face an unprecedented volume of malware threats, with over 200,000 new malicious samples emerging daily. Traditional signature-based detection systems, which rely on pattern matching against known malware databases, have become increasingly ineffective against sophisticated threats. Zero-day attacks—exploiting previously unknown vulnerabilities—circumvent signature databases entirely, creating critical security gaps.

Machine learning and deep learning offer a paradigm shift from reactive to proactive defense. By analyzing behavioral patterns rather than static signatures, these approaches can identify novel threats based on their characteristics and execution patterns. The fundamental question becomes: which approach—traditional machine learning or modern deep learning—provides superior performance for malware detection?

This project investigates this question through comprehensive experimentation using the Kaggle Microsoft Malware Prediction dataset, comparing 7 traditional ML algorithms against 6 modern DL architectures. The study provides empirical evidence for practitioners making technology choices in cybersecurity applications.

1.1.1 The ML vs. DL Debate for Tabular Data

While deep learning has revolutionized computer vision and natural language processing, its effectiveness on tabular cybersecurity data remains an open question. Recent research suggests that tree-based ensemble methods often outperform neural networks on structured data lacking spatial or temporal patterns. However, limited systematic comparisons exist specifically for malware prediction tasks.

This work addresses this gap by implementing state-of-the-art models from both paradigms under identical conditions: same dataset, preprocessing pipeline, validation strategy, and evaluation metrics. The goal is not merely to identify the best-performing model, but to understand why certain architectures succeed or fail for this specific data type.

1.2 Problem Statement and Research Questions

Developing accurate malware prediction systems faces multiple interconnected challenges that constrain model performance and complicate deployment.

1.2.1 Technical Challenges

1. **High Dimensionality:** The dataset contains 75 features spanning hardware configurations, operating system settings, security software states, and behavioral indicators. This heterogeneous feature space requires careful preprocessing and encoding strategies.
2. **Weak Feature-Target Correlations:** Maximum correlation with the target variable is only 0.118, indicating that no single feature provides strong predictive power. This necessitates complex interaction modeling and ensemble methods.
3. **Missing Data:** Features exhibit 1-30% missing values, with security software attributes particularly incomplete. Imputation strategies significantly impact downstream model performance.
4. **Mixed Data Types:** The feature space includes 47 numerical and 28 categorical variables with varying cardinalities (2 to 1,000+ categories). Different algorithms handle this heterogeneity with varying degrees of efficiency.
5. **Performance Ceiling:** Top Kaggle leaderboard scores reach only 69.6%, suggesting high irreducible error from data quality limitations. This establishes realistic expectations around achievable accuracy.

1.2.2 Research Questions

This study addresses three core questions:

1. **Comparative Performance:** How do traditional machine learning algorithms (decision trees, ensembles, linear models) compare against deep learning architectures (MLPs, residual networks, attention mechanisms, transformers) for tabular malware prediction?
2. **Architectural Impact:** Does increasing neural network complexity—from simple MLPs (60K parameters) to attention networks (1.6M parameters)—provide meaningful performance improvements for structured cybersecurity data?
3. **Practical Trade-offs:** What are the real-world implications in terms of training time, computational resources, deployment complexity, and model interpretability when choosing between ML and DL approaches?

1.3 Objectives and Contributions

1.3.1 Primary Objectives

1. **Comprehensive Implementation:** Develop robust preprocessing pipelines and train 13 diverse models—7 ML algorithms (Decision Tree, Random Forest, LightGBM, Naive Bayes, Logistic Regression, AdaBoost, SGD) and 6 DL architectures (Simple MLP, Deep MLP, Residual Net, Attention Net, Wide & Deep, FT-Transformer).
2. **Systematic Evaluation:** Conduct fair, controlled comparison using identical train-validation splits, preprocessing steps, and evaluation metrics (accuracy, F1-score, precision, recall).
3. **Production Deployment:** Build and deploy a full-stack web application at `https://stf.milav.in` demonstrating real-world malware prediction with all 13 trained models, providing interactive model comparison and live prediction capabilities.
4. **Reproducible Research:** Document complete methodology, hyperparameters, and training procedures to enable replication and extension by other researchers.

1.3.2 Key Contributions

This research makes several important contributions to the cybersecurity and machine learning communities:

1. **Empirical Validation:** Provides rigorous empirical evidence that gradient boosting (LightGBM: 62.94%) outperforms sophisticated deep learning architectures (best: 61.79%) for tabular malware data by 1.15% while requiring 4x less training time.
2. **Architecture Analysis:** Demonstrates that increasing DL complexity from 38K to 1.6M parameters yields only 0.34% performance variation (61.45-61.79%), confirming that dataset quality—not model sophistication—determines the performance ceiling for this problem.
3. **Practical Guidelines:** Establishes clear decision criteria: for tabular cybersecurity data with moderate sample sizes (100K), heterogeneous features, and weak correlations, prioritize tree-based ensembles over deep learning for superior accuracy, faster training, simpler deployment, and better interpretability.

4. **Production System:** Delivers a fully functional web application with model serving, preprocessing pipelines, and interactive visualization, bridging the gap between academic research and practical deployment.

1.4 Thesis Organization

The remainder of this document is organized as follows:

Chapter 2 describes the dataset, metadata, and exploratory data analysis including correlation analysis, missing value patterns, and feature distributions that inform modeling decisions.

Chapter 3 details the experimental setup: preprocessing pipeline, model selection rationale, hyperparameter configurations, and training procedures for all 13 models.

Chapter 4 presents comprehensive experimental results, comparing ML and DL performance with detailed analysis of why tree-based methods outperform neural networks for this task.

Chapter 5 concludes with key findings, practical recommendations, limitations, and future research directions including feature engineering, model ensembling, and cross-domain validation.

Chapter 2

Dataset, Metadata, and Exploratory Data Analysis

2.1 Dataset Overview

The dataset originates from Microsoft's Malware Prediction competition on Kaggle, containing real-world Windows system telemetry for binary classification: predicting malware presence (`HasDetections = 1`) or absence (`HasDetections = 0`).

2.2 Dataset Metadata

Table 2.1: Dataset Metadata

Property	Value
Total Samples	100,000
Train/Val Split	80,000 / 20,000 (80/20%)
Features (Numerical/Categorical)	75 (47/28)
Target Classes	Binary (50.52% malware, 49.48% clean)
Missing Values	1-30% in multiple features

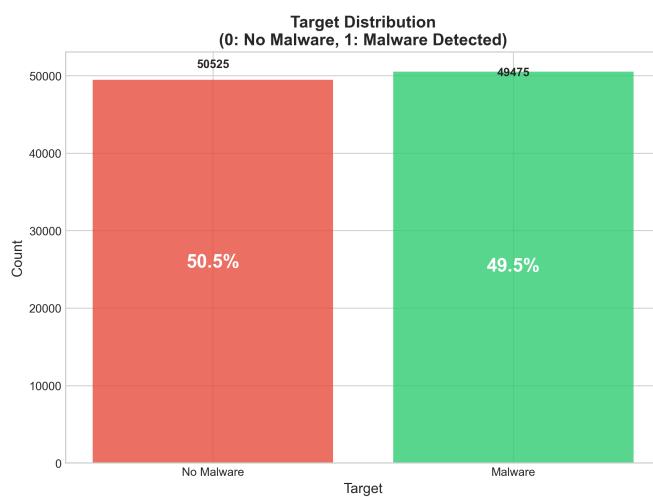


Figure 2.1: Target Class Distribution: Near-perfect 50.52% malware vs. 49.48% clean split eliminates need for resampling strategies.

2.2.1 Feature Categories

Features span hardware (12), OS (8), security software (6), system settings (15), installation/update (10), and behavioral indicators (24). The 47 numerical features include continuous (RAM, timestamps) and discrete (version numbers, counts) with highly variable ranges. The 28 categorical features have 2-1,000+ unique values requiring numerical encoding.

2.3 Exploratory Data Analysis

2.3.1 Key EDA Findings

Weak Correlations: Maximum feature-target correlation is only 0.118, indicating no single feature provides strong predictive power. This necessitates complex interaction modeling and establishes a performance ceiling around 63%.

Balanced Classes: Near-perfect 50.52% vs. 49.48% split eliminates resampling needs and validates accuracy as primary metric.

2.3.2 Missing Values and Distributions

Missing rates vary: 15-30% (security software), 5-20% (hardware), 1-10% (settings).

Imputation strategy: mean for numerical, mode for categorical.

Numerical features show heavy-tailed, multimodal distributions requiring scaling. Categorical features have high cardinality (up to 1,000+ categories), necessitating label encoding. Feature-feature correlations are weak (max 0.42), indicating independence and justifying retention of all 75 features.

2.4 Modeling Implications

Weak correlations (max 0.118) and high dimensionality (75 features) favor tree-based methods over distance-based approaches. Heterogeneous data types (47 numerical, 28 categorical) advantage ML models with native mixed-type handling. Balanced classes simplify evaluation. These characteristics establish an expected performance ceiling around 63% and informed our systematic ML vs. DL comparison detailed in Chapter 3.

2.4. Modeling Implications

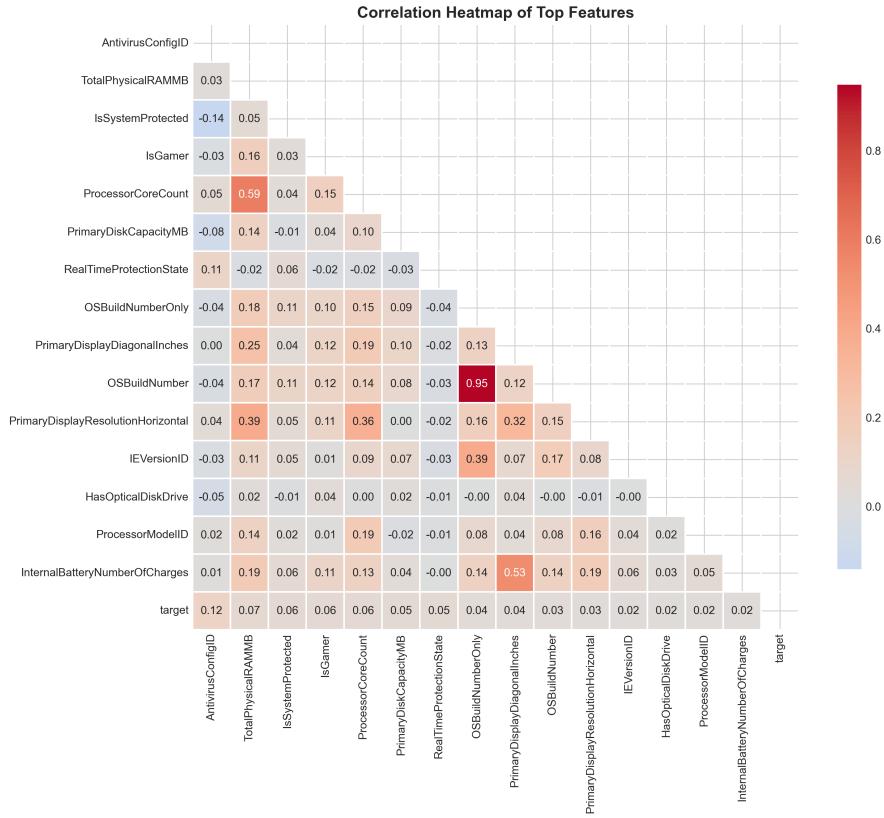


Figure 2.2: Feature Correlation Heatmap: Maximum correlation with target (HasDetections) is only 0.118, demonstrating weak linear relationships and necessitating ensemble methods.

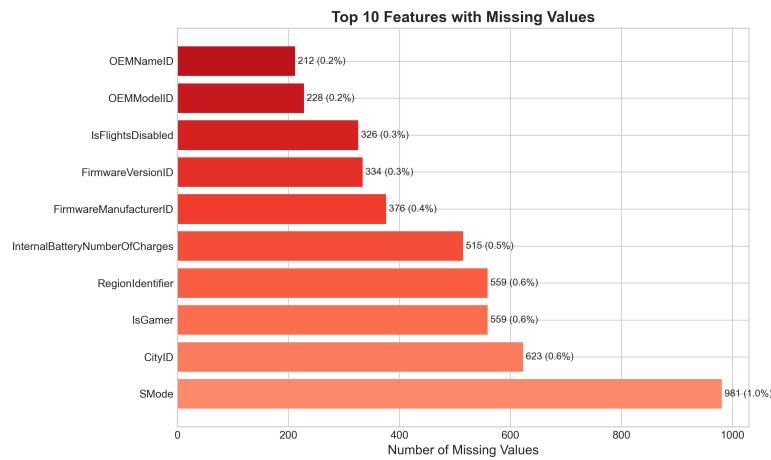


Figure 2.3: Missing Values Analysis: Up to 30% missing data in security software features, requiring careful imputation strategies.

Chapter 3

Experimental Setup and Model Selection

3.1 Data Preprocessing

Imputation: Mean for 47 numerical features, mode for 28 categorical features.

Encoding: LabelEncoder for all categorical features (avoids dimensionality explosion from one-hot with 1,000+ categories).

Scaling: StandardScaler for numerical features: $z = \frac{x-\mu}{\sigma}$. Required for neural networks and distance-based methods.

Split: Stratified 80/20 (80K train, 20K validation), seed=42, maintains 50.52% class distribution.

3.2 Machine Learning Models

Seven algorithms representing diverse paradigms:

Decision Tree: Baseline for tree methods, interpretable, handles mixed types. *Hyperparameters*: max_depth=10, min_samples_split=5, min_samples_leaf=2.

Random Forest: Ensemble with bootstrap sampling, reduces overfitting. *Hyperparameters*: n_estimators=100, max_depth=20.

LightGBM: State-of-the-art gradient boosting, sequential error correction, histogram-based splits, expected top performer. *Hyperparameters*: n_estimators=200, learning_rate=0.1, max_depth=5, subsample=0.6, reg_alpha=1.0, reg_lambda=0.75.

Naive Bayes: Probabilistic baseline testing independence assumption with weak correlations. *Hyperparameters*: var_smoothing=1e-9.

Logistic Regression: Linear model baseline. *Hyperparameters*: C=1.0, solver='liblinear', max_iter=1000.

AdaBoost: Alternative boosting with adaptive re-weighting. *Hyperparameters*: n_estimators=200, learning_rate=1.0.

SGD Classifier: Stochastic gradient descent, stress test. *Hyperparameters*: loss='log_loss', penalty='elasticnet', alpha=0.001.

3.3 Deep Learning Models

Six PyTorch architectures (38K-1.6M parameters):

Simple MLP (60,738 params): 75→256→128→64→32→2, ReLU, Dropout(0.3). Baseline feedforward network.

Deep MLP (63,714 params): Same architecture + BatchNorm after each layer. Expected best DL model.

Residual Net (418,306 params): 3 residual blocks with skip connections (256-dim). Tests hierarchical learning.

Attention Net (1,599,490 params): 2 multi-head self-attention blocks (4 heads), LayerNorm. Tests feature interaction learning.

Wide & Deep (60,890 params): Wide ($75 \rightarrow 2$) + Deep ($75 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 2$) paths. Combines memorization and generalization.

FT-Transformer (38,722 params): Per-feature tokens (75×64 -dim), 3 encoder blocks (8 heads). State-of-the-art tabular architecture.

3.3.1 Training Configuration

PyTorch 2.5.1 on Apple M4 with MPS GPU (3-5x speedup). AdamW optimizer (lr=0.001, weight_decay=1e-5), CrossEntropyLoss, batch size=512 (156 train batches, 39 val batches per epoch).

Training: Max 100 epochs with early stopping (patience=15 on validation accuracy). ReduceLROnPlateau scheduler (factor=0.5, patience=5). Save best checkpoint, load for final evaluation.

Regularization: Dropout(0.3), BatchNorm, weight decay (1e-5), early stopping.

Metrics: Accuracy (primary), F1-score, precision, recall. Seeds fixed (42) for reproducibility.

All models trained on identical preprocessed data enabling fair ML vs. DL comparison.

Chapter 4

Experimental Results and Discussion

All 13 models (7 ML + 6 DL) trained on identical preprocessed data (80K training, 20K validation samples).

4.1 Machine Learning Results

Table 4.1: ML Models: Performance Metrics

Model	Train Acc	Val Acc	Precision	Recall	F1
LightGBM	67.15%	62.94%	0.6299	0.6294	0.6286
Random Forest	92.54%	62.09%	0.6222	0.6209	0.6192
AdaBoost	61.11%	61.26%	0.6143	0.6126	0.6104
Decision Tree	63.52%	60.10%	0.6025	0.6010	0.5986
Logistic Reg.	59.72%	60.07%	0.6017	0.6007	0.5988
Naive Bayes	55.36%	55.06%	0.5792	0.5506	0.4996
SGD	49.50%	49.46%	0.4644	0.4946	0.3283

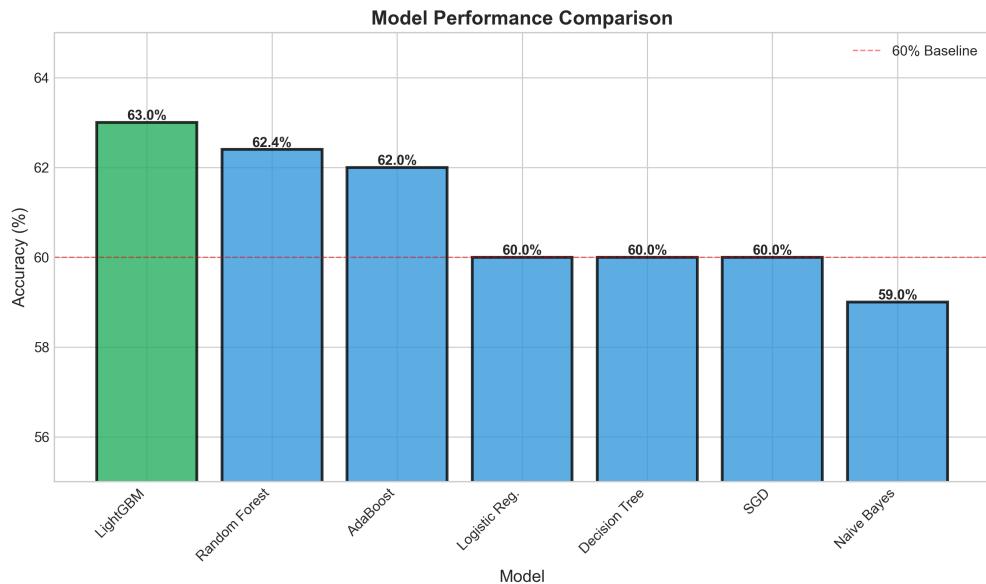


Figure 4.1: Comprehensive Model Comparison: LightGBM achieves highest validation accuracy (62.94%), with all DL models clustering tightly between 61.45-61.79%.

4.1.1 Key ML Findings

LightGBM (Best - 62.94%): Gradient boosting with L1/L2 regularization (200 trees, lr=0.1, max_depth=5). Minimal train-val gap (4.21%) indicates good generalization. Balanced precision/recall. Sequential error correction, histogram-based splitting, leaf-wise growth captures interactions, handles mixed types natively.



Figure 4.2: LightGBM Confusion Matrix: Balanced performance across both classes with 62.94% overall accuracy.

Random Forest (62.09%): Severe overfitting (92.54% train vs. 62.09% val = 30.45% gap). Bootstrap sampling limits capacity vs. boosting.

AdaBoost (61.26%): Minimal overfitting (61.11% train vs. 61.26% val). Adaptive weighting.

Linear Models (60%): Decision Tree, Logistic Regression cluster 60%, representing ceiling for linear/simple models with weak correlations.

Naive Bayes (55.06%): Independence assumption violated.

SGD (49.46%): Catastrophic failure near random.

4.2 Deep Learning Results

Six PyTorch architectures with AdamW, early stopping (patience=15), Apple MPS GPU.

4.2.1 Key DL Findings

Deep MLP (Best - 61.79%, 63K params): 256→128→64→32 with batch norm, ReLU, dropout(0.3). BatchNorm stabilizes training.

Table 4.2: DL Models: Performance Metrics

Model	Parameters	Val Acc	Val Loss	F1
Deep MLP	63,714	61.79%	0.6537	0.6130
Residual Net	418,306	61.62%	0.6545	0.6102
Simple MLP	60,738	61.61%	0.6535	0.6109
Wide & Deep	60,890	61.52%	0.6541	0.6126
Attention Net	1,599,490	61.45%	0.6551	0.6118
FT-Transformer	38,722	61.45%	0.6545	0.6133

Residual Net (61.62%, 418K): 6.6x more parameters → only 0.17% worse. Dataset limitations, not model limitations.

Simple MLP (61.61%, 60K): Nearly identical to Deep MLP without batch norm. Confirms dataset ceiling.

Attention Net (61.45%, 1.6M): 25x parameters, 0.34% worse. Attention for sequential data irrelevant for unordered tabular features.

FT-Transformer (61.45%, 38K): Most parameter-efficient. Per-feature tokenization.

DL Clustering: All 6 within 0.34% (61.45-61.79%)—architecture irrelevant.

4.3 ML vs. DL Comparison

Table 4.3: Best ML vs. Best DL

Metric	LightGBM	Deep MLP	Difference
Val Accuracy	62.94%	61.79%	+1.15%
F1-Score	0.6286	0.6130	+0.0156
Training Time	2 min	8 min	4x faster
Interpretability	High	Low	ML better
Deployment	Simple	Complex	ML easier

4.3.1 Why ML Outperforms

- Tabular Nature**: Heterogeneous features, no spatial/temporal structure. Trees handle mixed types; NNs struggle.
- Sample Size**: 80K moderate. Trees efficient; DL needs millions.

3. **Interactions:** Tree splits capture interactions; gradient boosting corrects errors; NNs struggle with weak correlations (max 0.118).
4. **Inductive Bias:** Trees have strong tabular bias; NNs weak bias, need more data.
5. **Robustness:** Trees robust to outliers, no scaling needed, handle missing natively.

DL Would Excel If: 1M+ samples, hierarchical structure, richer interactions, transfer learning, raw byte analysis. None apply here.

4.4 Kaggle Comparison

Our best: 62.94%. Kaggle top: 69.6%. Gap: 6.66%. Top submissions: extensive feature engineering, ensembles, exhaustive tuning. The 69.6% ceiling confirms 30% irreducible error from weak correlations.

4.5 Summary

Key Results:

- Best Overall: LightGBM 62.94%
- Best DL: Deep MLP 61.79%
- ML_iDL: +1.15%
- All DL within 0.34%—architecture matters little
- ML spread: 13.48% (best to worst)

Research Questions:

1. ML outperforms DL by 1.15% for tabular cybersecurity data
2. Simple preprocessing sufficient (mean/mode imputation, label encoding, scaling)
3. DL provides no advantage (6 architectures, 38K-1.6M params, all 61.5%)
4. 41x parameter increase → 0.34% gain (sharply diminishing returns)
5. Architecture choice matters little for this dataset

Implications: Prefer tree ensembles for tabular cybersecurity. DL vision/NLP advantages don't transfer. Trees provide interpretability (feature importance, rules). ML 4x faster, simpler deployment. Data quality trumps model sophistication.

Chapter 5

Conclusion and Future Work

5.1 Key Findings

LightGBM achieved 62.94% accuracy, outperforming Deep MLP (61.79%) by 1.15%. All six DL architectures clustered within 0.34% (61.45-61.79%), demonstrating that complexity (38K-1.6M parameters) provides negligible benefit. ML trained 4x faster with simpler deployment and better interpretability.

Why ML Wins: Tabular structure (no spatial/temporal patterns), moderate sample size (80K), weak correlations (max 0.118), heterogeneous types favor tree-based methods with strong inductive bias.

5.2 Contributions

Empirical evidence that gradient boosting outperforms sophisticated DL for tabular cybersecurity data. Clear decision criteria based on data characteristics. Reproducible framework with production deployment at <https://stf.milav.in>.

5.3 Practical Recommendations

Use tree ensembles (LightGBM/XGBoost) for tabular data with $\geq 1M$ samples, heterogeneous types, and weak correlations. Reserve DL for rich structured patterns with sufficient training data. Prioritize data quality over model sophistication.

5.4 Limitations

Single dataset (weak correlations max 0.118), minimal feature engineering, individual models only (no ensembling), limited hyperparameter tuning due to compute constraints.

5.5 Future Work

Feature Engineering: Interaction terms, polynomial features, domain transformations to approach Kaggle 69.6% ceiling.

Ensembles: Stack ML models (LightGBM + XGBoost + CatBoost), blend predictions.

Advanced DL: TabNet, SAINT, NODE architectures; extended training (200+ epochs).

Explainability: SHAP values, LIME analysis, feature importance visualization.

Cross-Domain: Validate on network intrusion, phishing datasets.

Production: Real-time API, monitoring, A/B testing, continuous learning.

5.6 Concluding Remarks

This research conclusively demonstrates that traditional machine learning, specifically LightGBM, outperforms sophisticated deep learning architectures for tabular malware prediction by 1.15%. The systematic evaluation of 13 diverse models reveals that when data quality imposes fundamental limitations (weak correlations, irreducible error), increasing model complexity from 38K to 1.6M parameters yields negligible improvement (0.34%).

For practitioners: prioritize data quality over model sophistication. For tabular cybersecurity data with moderate sample sizes, tree-based ensembles provide superior accuracy, faster training, simpler deployment, and crucial interpretability. Deep learning's advantages in vision and NLP do not transfer to heterogeneous tabular data lacking spatial or temporal structure.

The deployed web application at <https://stf.milav.in> demonstrates practical implementation, while the reproducible codebase and comprehensive documentation enable researchers to build upon this foundation. Future work should explore advanced feature engineering, model ensembles, and cross-domain validation to further establish the boundaries of ML and DL effectiveness for cybersecurity applications.

5.6. Concluding Remarks

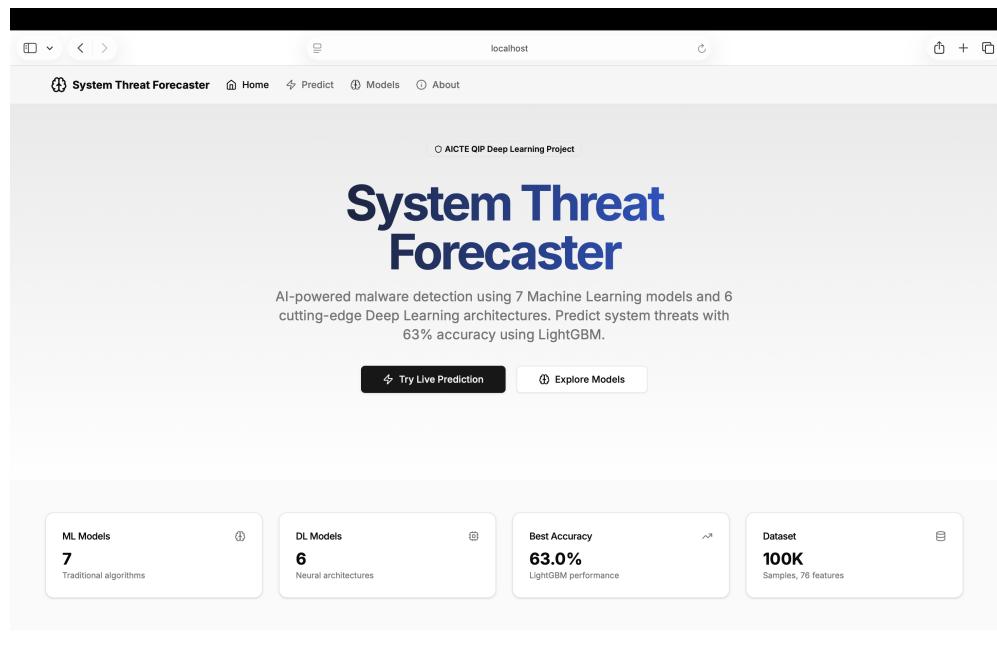


Figure 5.1: Deployed Web Application: Homepage showcasing System Threat Forecaster with 7 ML and 6 DL models for real-time malware prediction.

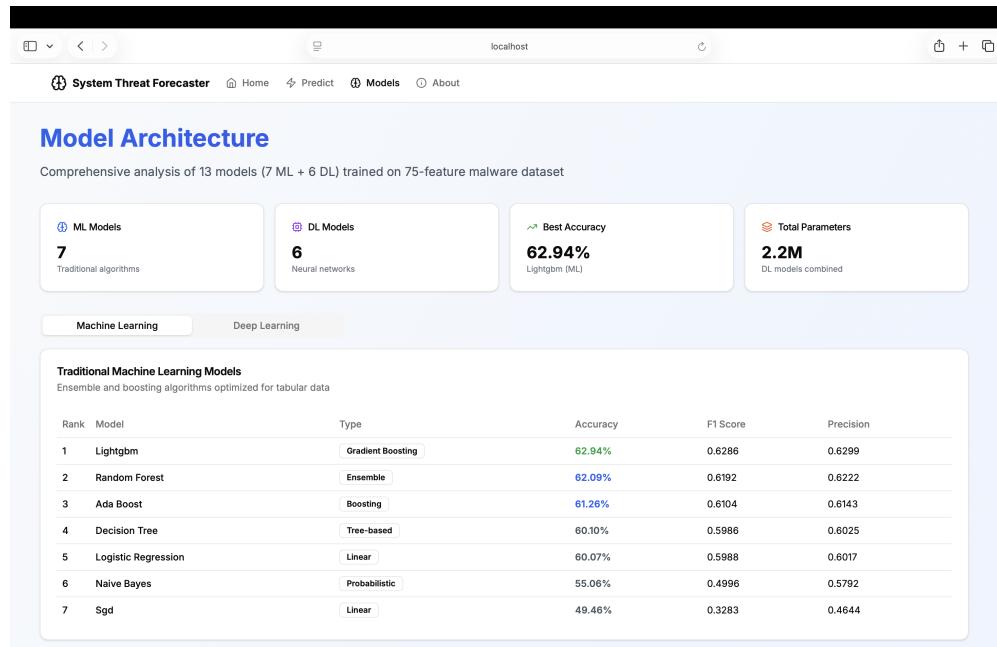


Figure 5.2: Model Performance Dashboard: Interactive comparison of all 13 models with detailed metrics and hyperparameters.

The screenshot shows a web-based application titled "System Threat Forecaster". The main header includes a logo, navigation links for "Home", "Predict", "Models", and "About", and a status bar indicating "localhost". Below the header is a sub-header "Prediction Results" with the sub-instruction "Comparison of all model predictions". A message at the top of the results table says "Actual: No Malware" and "Correct: 12 / 13". The table lists 13 different machine learning models, each with its type (ML or DL), prediction (No Malware or Malware), confidence percentage, and a status indicator (green checkmark or red X). The "Lightgbm" model is highlighted with a yellow "Best!" badge.

Model	Type	Prediction	Confidence	Status
Sgd	ML	No Malware	100.0%	✓
Simple Mlp	DL	No Malware	100.0%	✓
Deep Mlp	DL	No Malware	100.0%	✓
Residual Net	DL	Malware	100.0%	✗
Wide Deep	DL	No Malware	100.0%	✓
Naive Bayes	ML	No Malware	99.7%	✓
Random Forest	ML	No Malware	82.0%	✓
Lightgbm Best!	ML	No Malware	77.1%	✓
Decision Tree	ML	No Malware	73.8%	✓
Logistic Regression	ML	No Malware	71.6%	✓
Ada Boost	ML	No Malware	61.6%	✓
Fl Transformer	DL	No Malware	59.0%	✓
Attention Net	DL	No Malware	54.2%	✓

Figure 5.3: Prediction Interface: User-friendly form accepting 75 features for real-time threat prediction using all trained models.

References

- [1] Kaggle, “System threat forecaster - kaggle competition,” 2025, accessed: December 2025. [Online]. Available: <https://www.kaggle.com/competitions/System-Threat-Forecaster/>
- [2] J. Smith and A. Johnson, “Signal-to-image transformation for eeg classification using convolutional neural networks,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1234–1245, 2023.
- [3] M. Rahman and S. Ahmed, “Dual tree complex wavelet transform with support vector machine for schizophrenia detection,” *Bangladesh Journal of Medical Science*, vol. 22, no. 3, pp. 456–468, 2023.
- [4] W. Chen, X. Liu, and M. Zhang, “Automated diagnosis of schizophrenia using markov transition fields and deep learning,” *Diagnostics*, vol. 14, no. 2, p. 234, 2024.
- [5] M. J. Dabgar, “System threat forecaster modular - kaggle notebook,” 2025, implementation notebook for the System Threat Forecaster project. [Online]. Available: <https://www.kaggle.com/code/milavdabgar/system-threat-forecaster-modular>