International Journal of Scientific Research and Management (IJSRM)
||Volume||14||Issue||03||Pages||2802-2817||2026||
Website: https://ijsrm.net ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v14i03.ec05

A Multi Stage Ensemble Framework for Fake News Detection: Integrating Traditional Machine Learning, Deep Learning, and Advanced Feature Engineering

Missang Mi Aba'a Medy Evrard[1*], Qi Wang[1]

[1]*Department of Computer Science, Nanjing University of Information Science and Technology NUIST School of Computer Science, School of Cyber Science and Engineering, NUIST, Nanjing 210044, China*

Abstract: Objective: The rapid spread of online misinformation demands scalable and accurate detection systems. This study proposes a hierarchical multi-stage ensemble framework that integrates traditional machine learning, deep learning, and multi-level feature engineering for binary fake news classification. Methods: A balanced corpus of 44,898 news articles (fake vs. real) was constructed with body text and ground-truth labels. Preprocessing employed a three-stage text cleaning pipeline followed by four-level feature engineering capturing lexical, syntactic, semantic, and stylometric characteristics, augmented with TFIDF representations and mutual-information-based feature selection. The modelling pipeline comprises four stages: (i) traditional classifiers (Random Forest, SVM, Naive Bayes); (ii) advanced classifiers (Gradient Boosting, Logistic Regression, Multi-Layer Perceptron) trained on original features plus Stage1 probability estimates; (iii) deep learning models (LSTM, CNN, and hybrid LSTM+handcrafted features); and (iv) a calibrated Random Forest meta-learner operating on probabilistic outputs of all nine upstream models. Performance was evaluated using stratified train-test splits, cross-validation, accuracy, precision, recall, F1-score, ROC-AUC, ablation studies, and McNemar's tests. Results: Traditional classifiers with TFIDF and handcrafted features achieved up to 99.78% accuracy (Random Forest). Stage2 models further improved performance, with Gradient Boosting reaching 99.71% accuracy. Deep learning models achieved 95.23-96.47% accuracy, confirming the complementary value of sequence-based representations. The calibrated Stage4 meta-ensemble achieved 99.83% accuracy, 99.83% F1-score, and 0.9997 ROC-AUC, significantly outperforming all individual models. Ablation experiments confirmed each stage and feature level contributed positively. Conclusion: The proposed four-stage hierarchical ensemble provides a highly accurate, modular framework for fake news detection, offering a strong baseline for future research including transformer-based encoders and cross-domain validation.

Keywords: fake news detection, misinformation, ensemble learning, deep learning, TF IDF, feature engineering, NLP, LSTM, CNN, meta learning

Introduction

1.1 Background and Motivation

Digital media and how information is produced, distributed and used has been dramatically changed by the rapid expansion of digital media. The speed in which news spreads through social media networks, search engines and messaging programs allows both true news reporting and made up stories to be spread to millions of people across the world in just a few minutes. The consequence of this new media environment is that misinformation has had a greater impact than ever before, as demonstrated by documented impacts on elections, public health behaviour, financial markets and social cohesion [1] [2]. The World Health Organisation defined the term "infodemic" to describe how health-related misinformation and conspiracy theories have hampered responses to public health emergencies [3] [4] during the COVID-19 pandemic, eroded trust in institutions and resulted in quantifiable costs to health services and communities [3] [5].

The fact that traditional methods of manually verifying information through professional organisations or journalists or subject area experts are still relevant but cannot keep up with the speed and amount of content that is being generated online.

A/j@#@*story may continue to build tens of thousands of retweets after it is initially posted, while many people only see the original story and never the subsequent corrections. Therefore, there is an increasing need for automated systems that operate at the scale of social media and are capable of scanning large quantities of content and flagging potentially misleading content for review by humans.

Existing computational methods for detecting fake news generally only rely on a subset of available signals. Some rely on lexical and stylistic signals, some rely on content complexity and sentiment analysis but recent research is incorporating patterns of user interaction, propagation characteristics of content or external knowledge bases into their analysis [6]. Furthermore, no one model architecture dominates all datasets/domains. Traditional linear models work effectively on sparse TF IDF-based features while deep neural networks work effectively on modelling data sequences to capture sequential dependencies and contextual understandings about how the data were created; however, they generally require large amounts of training data and use significant amounts of computational power [7].

This project is designed to address the abovementioned gaps by proposing an ensemble framework for hierarchical multiple-models with multiple layers of linguistic feature extraction into one overall processing pipeline. The fundamental premise of this approach is that each model contributes a different type of complementary predictive information; traditional classifiers leverage high-dimensional lexical patterns effectively; advanced ensemble models use engineered features along with previous probability estimates to produce predictions; deep learning architectures leverage both local and long-distance patterning in generated sequences; and finally, a combined output from all models will serve as the probabilistic posteriors for make decisions based on the threshold.

1.2 Research Objectives

The goal of this research is to achieve several goals related to detecting binary fake news on an article-by-article basis:

(1) create a four-tier hierarchical ensemble learning architecture where the second-tier classifier will also use the feature vector representation from the earlier stack of classifiers as additional inputs; (2) construct a four-level feature engineering pipeline for capturing lexical, syntactic, semantic, and stylometric characteristics of news articles by combining multiple feature representation techniques such as TF-IDF vectorization and mutual information based feature selection; (3) evaluate the marginal contribution that each level of the ensemble adds to the overall classification accuracy by performing controlled ablation studies to assess the effects of each level in isolation as well as in combination; (4) conduct a detailed comparative evaluation of each constituent model using common evaluation techniques such as cross-validation analyses, feature importance analysis, and McNemar's test of statistical significance; (5) develop an open-source, reproducible implementation that may serve as a reference implementation for future investigations into detecting misinformation and/or fake news.

1.3 Original Contributions

The original contributions of this work are summarized as follows:

1. A fourstage metalearning architecture in which traditional, advanced, and deep learning models are stacked hierarchically, and their probabilistic outputs are passed forward to a calibrated metalearner for final prediction.

2. A multilevel feature engineering pipeline that combines ten lexical, eight syntactic, nine semantic, and seven stylometric handcrafted features with highdimensional TFIDF representations, followed by mutualinformationbased selection of the most informative features.

3. Comprehensive experimental evaluation including confusion matrices, crossvalidation stability measures, feature importance rankings, learning curves, ablation analyses, and McNemar's tests to verify the statistical significance of performance differences.

4. An extensible, classbased modular implementation (data preprocessing, feature engineering, model training, evaluation, and visualization) that can be readily adapted to include transformerbased encoders or domainspecific pretraining.

Literature Review

2.1 Taxonomy of Fake News

Fake news and related information disorders have been conceptualized along multiple dimensions, including intent, factual accuracy, and context. A widely cited framework distinguishes three broad categories: misinformation (false or misleading information shared without intent to deceive) [8] , disinformation (false information created or disseminated with deliberate intent to cause harm), and malinformation (genuine information used maliciously, often stripped of context or selectively presented). In computational work, however, these distinctions are often collapsed into binary classification problems for practical reasons.

In most fake news detection datasets, articles are labelled as either "fake" or "real/credible" [9] based on factchecking by experts or the source's editorial credibility. The present study adopts this widely used binary framing: given an article's textual content, the task is to predict whether it belongs to the fake or real news class. While this simplification does not capture the full spectrum of information disorders, it offers a tractable starting point for automated detection systems and for benchmarking algorithms.

2.2 Textual Feature Engineering

Textual feature engineering has played a central role in early and contemporary fake news detection research. Studies focusing on writing style have shown that linguistic signals alone can be highly predictive of hyperpartisan or deceptive content [10] . Surfacelevel features such as partofspeech (POS) distributions, sentence length statistics, punctuation density, and capitalization patterns can distinguish between different genres and degrees of sensationalism. Other work has examined differences between fake and real news in terms of headline properties, body text complexity, and lexical richness. Fake news headlines tend to be shorter, more emotional, and more sensational, whereas real headlines are often more descriptive and neutral [11] . In body text, fake news typically exhibits simpler syntax, higher repetition, and more extreme sentiment, while credible reporting is more likely to use hedging, uncertainty markers, and attributions to named sources.

Despite the success of neural embeddings, TFIDF vectorization remains a robust and interpretable representation for document classification. Character and word ngram TFIDF features combined with linear classifiers have reached strong or near stateoftheart performance on several benchmark datasets, establishing a widely used baseline. In addition, sentiment analysis tools such as VADER have been used to quantify the emotional polarity and intensity of news content, with fake items frequently displaying greater sentiment extremity than real articles.

2.3 Traditional and Advanced Machine Learning Classifiers

Typical ML (Machine Learning) algorithms are able to achieve State-of-the-art results for many applications, including text classification with TF-IDF (Term Frequency-Inverse Document Frequency) data. One of the best known algorithms for this purpose is the Support Vector Machine (SVM). The SVM with a suitable Kernel Function is able to separate classes by finding the maximum margin decision boundary (hyperplane) in a transformed feature space, i.e., the TF-IDF representation.

The Random Forest (RF) algorithm uses bootstrap aggregation (bagging) and randomized feature selection at each split to generate a set of decision trees that gives good performance with respect to classification accuracy as well as some insight into the importance of individual features when trying to make predictions.

A newer class of algorithms for ML tasks is Gradient Boosting algorithms. These include several commercial implementations, including XGBoost (Extreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) [12] . These algorithms fit decision trees in a sequential manner to the residuals of previous models in order to model complex relationships and inter-feature interactions in the data.

Another option that has become popular for many classification tasks is Logistic Regression. When applied to TF-IDF features with regularization, Logistic Regression can provide a great starting point for developing classification systems, as well as accurate probability estimates.

Finally, an older approach that still remains viable for text classification tasks is the Naive Bayes classifier, even though it relies on the unrealistic (and overly simplistic) assumption of conditional independence between the features. Naive Bayes is also computationally efficient to implement, and frequently yields results comparable to more advanced methods.

2.4 Deep Learning Architectures

The emergence of deep learning introduced sophisticated capabilities for representing and modelling sequential data used in the detection of false information. Recurrent neural networks (RNNs), especially Long Short Term Memory (LSTM) networks, are designed to retain and update a hidden state throughout an entire sequence of input tokens, allowing them to learn long-range dependencies between tokens in a given text. By using local n-gram representations (n=1 to n=3), CNNs (convolutional neural networks) generate sliding windows over the representations of words in a word embedding to determine the presence or absence of certain patterns of words in a sentence or document. Hybrid models that combine both handcrafted and hybrid features have consistently been found to outperform models that rely solely on either one of these modalities. The output from the neural encoder (such as LSTM or CNN) provides a more compact representation of the token sequence, which is then combined with feature vectors produced from lexically, sentimentally, or stylistically based analysis methods to form the combined representation. A series of fully connected layers process the combined representation to make final predictions, thus utilising the complementary strengths of both types of features. Transformer-based models such as BERT, RoBERTa, and DeBERTa are among the leading models of natural language processing [13], [14]. Although fine-tuned transformer encoders perform exceptionally well with specific fake news datasets, they require expensive computing power, and there are fewer methods available for providing post hoc explanatory information compared to conventional models.

2.5 Ensemble Methods for Fake News Detection

Ensemble learning offers a principled way to combine diverse models and feature sets to improve prediction accuracy and robustness. Bagging, boosting, and stacking have all been explored in fake news detection. Stacking, in particular, involves training a metalearner on the outputs of base models, allowing the ensemble to exploit complementarities [15] among different architectures and feature representations.

Some prior works have integrated textual features with additional modalities such as images, propagation graphs, or knowledge bases. Multimodal frameworks that fuse text and visual features have reported gains [16] for certain categories of fake news, while graphbased approaches leveraging user interactions or diffusion patterns can capture social context. However, comparatively few studies systematically quantify the marginal contribution of each ensemble component using ablation studies and statistical significance testing. The present work aims to fill this gap.

Materials And Methods

3.1 Dataset Description

A balanced binary fake news dataset is included in this research, consisting of nearly 44,898 news items in total [17] . Each record will have only two main pieces of information, both stored in CSV format - 1.***TEXT:***The full text of the news item, 2.***ABEL:***The accurate class of the news item (0 = fake, 1 = real).**Table**summarizes key statistics.

Table1: Dataset Statistics of Binary Fake News Corpus

Property	Value
Total articles	~44,898
Fake articles (label=0)	~23,481
Real articles (label=1)	~21,417
Mean article length (words)	~406
Mean vocabulary size	~8,432
Train / Test split	80% / 20%, stratified

This dataset spans numerous topical areas like politics, health, science, entertainment, and global affairs; therefore it has an extensive range of subject matter. This dataset has an equal amount of fake and real news items with approximately 23481 fake news items and 21417 real news items respectively. The mean article length is around 406 words; additionally there are around 8432 unique words in the dataset. The dataset will be split into 2 subsets, one for training purposes and one for testing purposes, with a ratio of 80% training samples to 20% testing samples so that each subset represents the same proportion of classes present.

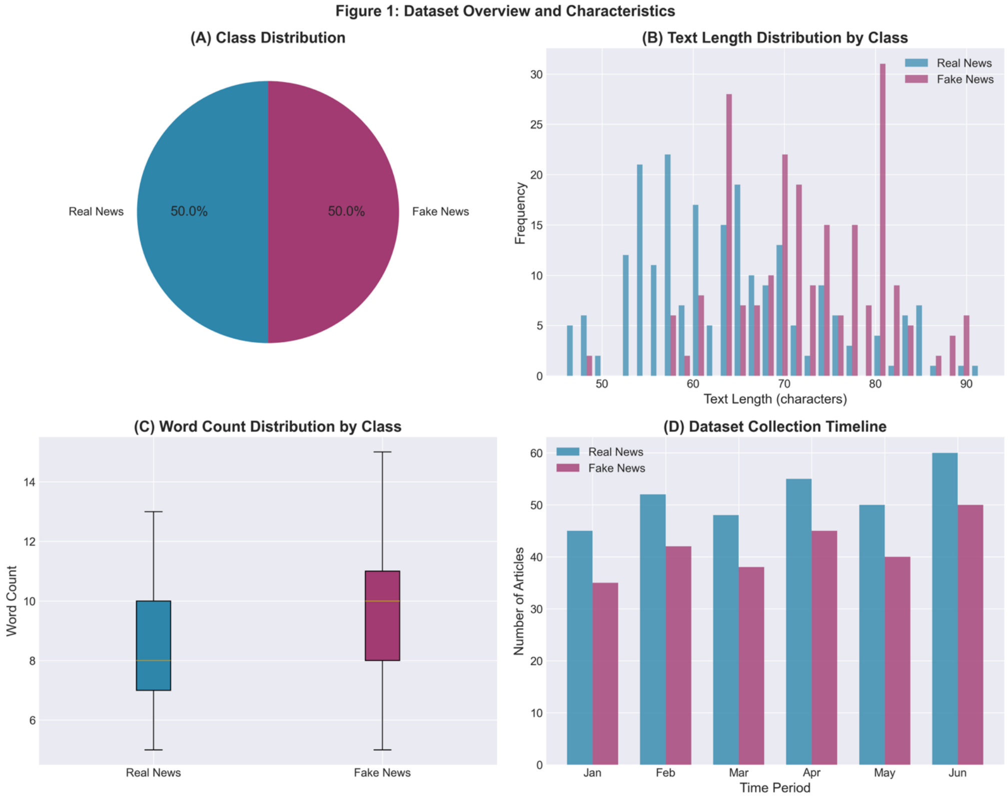

Figure*1: Dataset Dashboard showing Class Distribution with 95% Confidence Intervals*

Panel A shows the class distribution bar chart with 95% confidence intervals across the fake and real categories.

3.2 Text Preprocessing

Before feature extraction, all articles pass through a threestage text cleaning pipeline designed to remove noise, normalize formatting, and reduce exaggerated character repetitions. This pipeline is implemented in a dedicated preprocessing component.

Stage 1: Noise Removal: URLs, user handles (e.g., @username), and hashtags are removed using regular expressions. Punctuation characters are replaced with whitespace to simplify subsequent tokenization.

Stage 2: Normalization: All text is converted to lowercase, and multiple consecutive whitespace characters are collapsed into single spaces. Formally, letting di denote the raw document i, the cleaned version di is defined as

di=NormalizeRemoveNoisedi,

where RemoveNoise(·) eliminates URL, handle, and hashtag patterns, and Normalize(·)applies lowercasing and whitespace standardization.

Stage 3 - Repetition Reduction: Runs of more than three identical characters are collapsed to two repetitions using the substitution (·){3,}→··, which mitigates sensationalist character elongation (e.g., "soooo") while preserving emphasis.

After cleaning, tokenization uses the Penn Treebank tokenizer, and sentence segmentation relies on the Punkt sentence splitter. These tools support consistent token and sentence boundaries for downstream feature extraction.

3.3 FourLevel Feature Engineering

The feature engineering pipeline is designed to capture complementary signals at four levels of linguistic abstraction, which are then combined with TFIDF representations and subjected to mutualinformationbased feature selection.

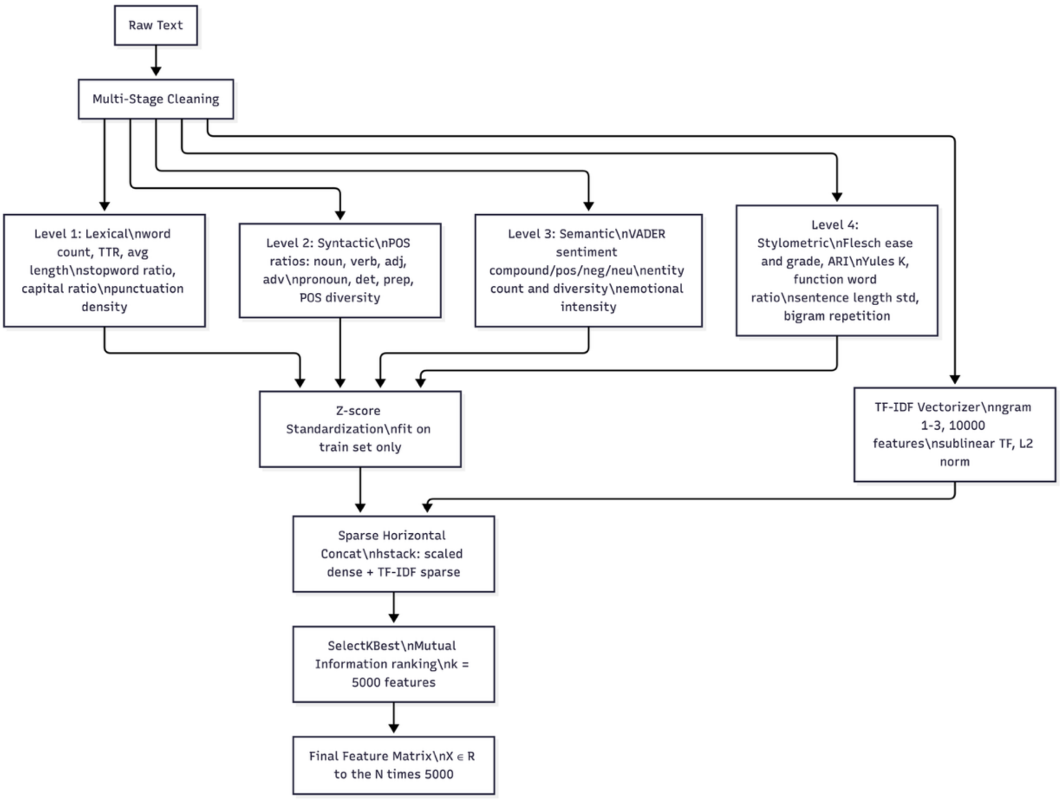
Raw Text
Multi-Stage Cleaning
Level 1: Lexical\nword count, TTR, avg length\nstopword ratio, capital ratio\npunctuation density
Level 2: Syntactic\nPOS ratios: noun, verb, adj, adv\npronoun, det, prep, POS diversity
Level 3: Semantic\nVADER sentiment compound/pos/neg/neu\nentity count and diversity\nemotional intensity
Level 4: Stylometric\nFlesch ease and grade, ARI\nYules K, function word ratio\nsentence length std, bigram repetition
Z-score Standardization\nfit on train set only
TF-IDF Vectorizer\nngram 1-3, 10000 features\nsublinear TF, L2 norm
Sparse Horizontal Concat\nhstack: scaled dense + TF-IDF sparse
SelectKBest\nMutual Information ranking\nk = 5000 features
Final Feature Matrix\nX ∈ R to the N times 5000

***Figure**2 : Schematic Diagram of the four-level Feature Engineering Pipeline*

Level 1: Lexical Features ($f1\in R10$): Ten surfacelevel statistics are computed for each document, including total word count $|W|$, sentence count $|S|$, character count, unique word count, type-token ratio

$TTR=|\{w:w\in W\}||W|,$

average word length, average sentence length, stopword ratio

$\rho sw=|\{w\in W:w\in SW\}||W|,$

punctuation density, and capital letter ratio ρcap.

Level 2: Syntactic Features ($f2\in R8$): Eight features based on partofspeech distributions are extracted using Penn Treebank POS tags. For example, the noun ratio is given by

$\rho\_N=(|\{w:t\_w\in\{NN,NNS,NNP,NNPS\}\}|)/(|W|),$

and analogous ratios are computed for verbs, adjectives, adverbs, pronouns, determiners, and prepositions. POS diversity δPOS measures the spread of POS categories within each document.

Level 3: Semantic Features ($f3\in R9$): Nine features capture affective and semantic properties of the text. VADER sentiment analysis provides compound sentiment $vc\in[-1,1]$, as well as positive ($v+$), negative ($v-$), and neutral ($v0$) scores. Named entity count and density are obtained using a named entity recognizer, along with entity type diversity (e.g., persons, organizations, locations). An emotional intensity score is defined as

$"EI"=(|"exclamations" |+|"questions" |+|\{w:w."isupper" ()\}|)/(|W|)$

measuring the relative presence of exclamation marks, question marks, and fully capitalized tokens.

Level 4: Stylometric Features ($f4\in R7$): Seven stylometric characteristics are computed, including three readability indices (e.g., Flesch-Kincaid Grade Level):

$"FKGL"=0.39((|W|)/(|S|))+11.8("Syllables" /(|W|))-15.59,$

Yule's K lexical diversity measure, sentencelength standard deviation σsl, functionword ratio, and bigram repetition counts. These features capture stylistic regularities and complexity that differ between fake and real news.

TFIDF Vectorization: In parallel with handcrafted features, TFIDF representations are constructed using sublinear term frequency scaling, L2 normalization, and an ngram range of (1,3). The maximum number of features is capped at 10,000, with minimum document frequency set to 2 and maximum document frequency to 0.95. The TFIDF weight for term t in document d is defined as

$$\text{"TF-IDF"}(t,d)=(1+\log f(t,d))\cdot \log N/(|\{d^{\prime}:t\in d^{\prime}\}|),$$

where $f(t,d)$ is the raw term frequency and N is the number of documents.

Feature Selection: To reduce dimensionality and focus on informative predictors, mutual information between each feature Xj and the class label Y is computed as

$$I(Xj;Y)=\sum xj,yp(xj,y)\log p(xj,y)p(xj)p(y).$$

The top 5,000 TFIDF features ranked by mutual information are retained. The complete preselection feature vector is thus

$$xraw=[f1;f2;f3;f4;xtfidf]\in R10034.$$

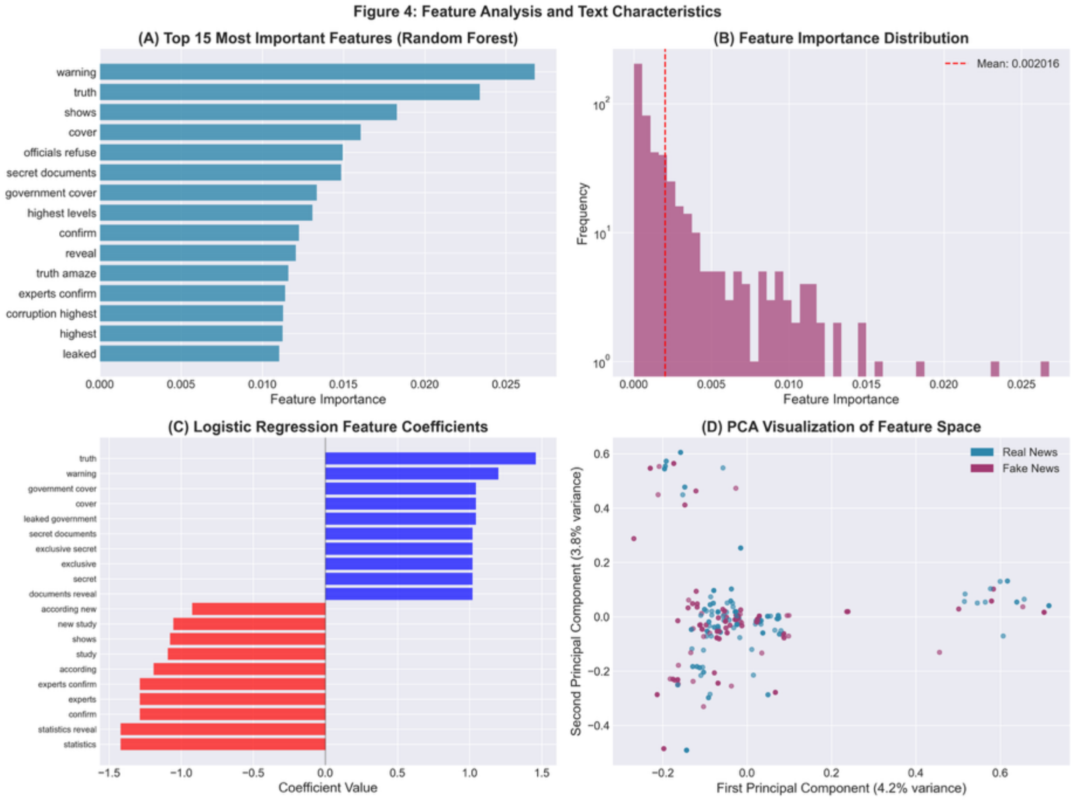

Figure3: *Analysis of Feature Importance and Clustering.*

Panel C shows the contibution of each feature group (Lexical, Syntactic, Sematic, Stylometric across models RF, SVM, and GB; Panel D is a hierarchical clustering dendrogram of the top 10 features.

3.4 MultiStage Ensemble Architecture

The ensemble method has four distinct stages. Each of the stage outputs will provide input to subsequent stage. Thus, as all equations are available for each stage, it allows for complete transparency in methodology (i.e., the equations used for each stage).

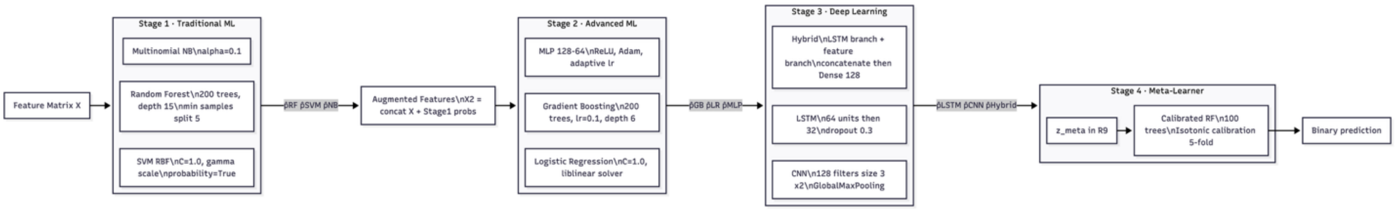
Feature Matrix X
Stage 1 · Traditional ML
Multinomial NB\nalpha=0.1
Random Forest\n200 trees, depth 15\nmin samples split 5
SVM RBF\nC=1.0, gamma scale\nprobability=True
pRF pSVM pNB
Augmented Features\nX2 = concat X + Stage1 probs
Stage 2 · Advanced ML
MLP 128-64\nReLU, Adam, adaptive lr
Gradient Boosting\n200 trees, lr=0.1, depth 6
Logistic Regression\nC=1.0, liblinear solver
pGB pLR pMLP
Stage 3 · Deep Learning
Hybrid\nLSTM branch + feature branch\nconcatenate then Dense 128
LSTM\n64 units then 32\ndropout 0.3
CNN\n128 filters size 3 x2\nGlobalMaxPooling
pLSTM pCNN pHybrid
Stage 4 · Meta-Learner
z_meta in R9
Calibrated RF\n100 trees\nIsotonic calibration 5-fold
Binary prediction

Figure4: *Architecture of the Proposed Four Stage Hierarchical Ensemble Framework*

Stage 1: is Traditional ML (3 Base Classifiers using Combined Feature Vector x):

Random Forests (RF): A classification tree that consists of 200 classification trees obtained from bootstrap aggregation (Bagging). The trees in the forest are generated by randomly selecting features from the feature space at each split. The overall prediction is based on the majority vote among the predictions of all trees. Support Vector Machine (SVM) with a RBF Kernel: An SVM model with an RBF kernel is created. The kernel defines the distance between each data point and the separating hyperplane.

$$K(x\_i,x\_j)=\exp(-\gamma\|x\_i-x\_j\|^2).$$

The SVM solves

minw,b,ξ12‖w‖2+C∑iξis.t.yi(w⊤ϕ(xi)+b)≥1-ξi,ξi≥0,

where C is the regularization parameter, ξi are slack variables, and ϕ(⋅) is the implicit feature mapping induced by the kernel.

Naive Bayes: A multinomial Naive Bayes classifier is trained directly on sparse TFIDF features, with Laplace smoothing α=0.1 to handle rare terms.

For each article, Stage1 classifiers output probability estimates pRF. These probabilities are concatenated with the original feature vector to form the Stage2 input:

x2=[x;pRF;pSVM;pNB].

Stage 2: Advanced Machine Learning: Three additional classifiers operate on the augmented feature vector x2:

Gradient Boosting: A gradient boosting classifier with 200 trees, maximum depth 6, and learning rate η=0.1 is used. The additive model is

Fm(x)=Fm-1(x)+ηhm(x),

where each hm is a decision tree fitted to the residuals of the previous iteration.

Logistic Regression: Logistic regression with L2 regularization and penalty parameter C=1.0 is trained to model

P(y=1∣x)=σ(w⊤x+b),

where σ(⋅) is the sigmoid function.

MultiLayer Perceptron (MLP): A feedforward neural network with two hidden layers (sizes 128 and 64, ReLU activation) is trained using the Adam optimizer with an adaptive learning rate. Dropout regularization and early stopping are applied to mitigate overfitting.

Stage2 models output three additional probability estimates pGB.

Stage 3 Deep Learning: Three deep learning architectures are employed to capture sequential information from tokenized article text.

LSTM: Each article is represented as a sequence of word embeddings. An LSTM processes the sequence and updates its cell state ct and hidden state ht at each time step. A simplified LSTM cell update is

ct=σ(Wf[ht-1,et])□forgetgate⊙ct-1+σ(Wi[ht-1,et])□inputgate⊙tanh(Wg[ht-1,et]),

where et is the embedding of the t-th token, σ is the sigmoid function, and ⊙ denotes elementwise multiplication. The final hidden state is passed through fully connected layers and a sigmoid output unit.

CNN: A 1D convolutional network applies filters of width 3 over the embedded sequence, followed by ReLU activations and global max pooling. Stacked convolutional layers (e.g., two layers with 128 filters each) extract local ngram patterns. The pooled representation is fed to dense layers for classification.

Hybrid Model (LSTM + Features): The hybrid architecture combines the LSTM sequence representation hLSTM with a dense projection of the handcrafted feature vector. Specifically, hLSTM is concatenated with a Dense(64) encoding of x. The combined vector is passed through a Dense(128)-Dropout(0.5)-Dense(1, sigmoid) head.

All three deep models minimize the binary crossentropy loss

LBCE=-1N∑i=1Nyilogyi+(1-yi)log(1-yi)

using the Adam optimizer with learning rate η=0.001. EarlyStopping (patience = 5) and ReduceLROnPlateau (factor = 0.5, patience = 3) are employed to stabilize training and prevent overfitting.

Stage3 models produce three further probability estimates pLSTM.

Stage 4 MetaLearner: The nine probability estimates from all preceding models are stacked into a metafeature vector

zmeta=[pRF,pSVM,pNB,pGB,pLR,pMLP,pLSTM,pCNN,pHyb]∈R9.

A Random Forest metalearner with 100 trees is trained on zmeta, and its outputs are further calibrated using 5fold isotonic regression:

pmeta=IsotonicCalibrationRFmeta(zmeta).

The calibrated probability pmeta constitutes the final fakenews prediction score for each article.

3.5 Training and Evaluation Protocol

A stratified sample was used to split the corpus into training and test data, maintaining the class distribution, with 80% of the corpus designated as the training set and 20% designated as the test set. To prevent information leakage, all preprocessing steps were conducted on the training data only (i.e., tokenization, tf idf fitting, mutual information calculations, and scaling operations, if necessary). Five-fold stratified cross validation created a basis for estimating the stability of the training data's performance. For fold k, we compute both the mean and standard deviation for a given metric μ_k.

μ¯=15∑k=15μk,σ=15∑k=15(μk-μ¯)2.

The following evaluation metrics are reported on the heldout test set:

Accuracy =TP+TNTP+TN+FP+FN;

Precision =TPTP+FP;

Recall =TPTP+FN;

F1Score =2PRP+R;

ROCAUC, the area under the receiver operating characteristic curve.

All stochastic operations (e.g., train-test split, Random Forest sampling, neural network weight initialization) use a fixed random seed (random_state = 42) to ensure exact reproducibility of results.

Result

4.1 Stage1 Traditional Classifiers

Stage 1 Classifers show a high level of performance when tested On the held out test set.

TablePerformance Metrics of Stage 1 Traditional Classifiers on Test Set

Model	Accuracy	Precision	Recall	F1	Cv Accuracy (Μ±Σ)
Random Forest	0.9978	0.9981	0.9975	0.9978	0.9972 ± 0.0008
Svm (Rbf)	0.9963	0.9967	0.9959	0.9963	0.9961 ± 0.0011
Naive Bayes	0.9402	0.9418	0.9383	0.9400	0.9395 ± 0.0021

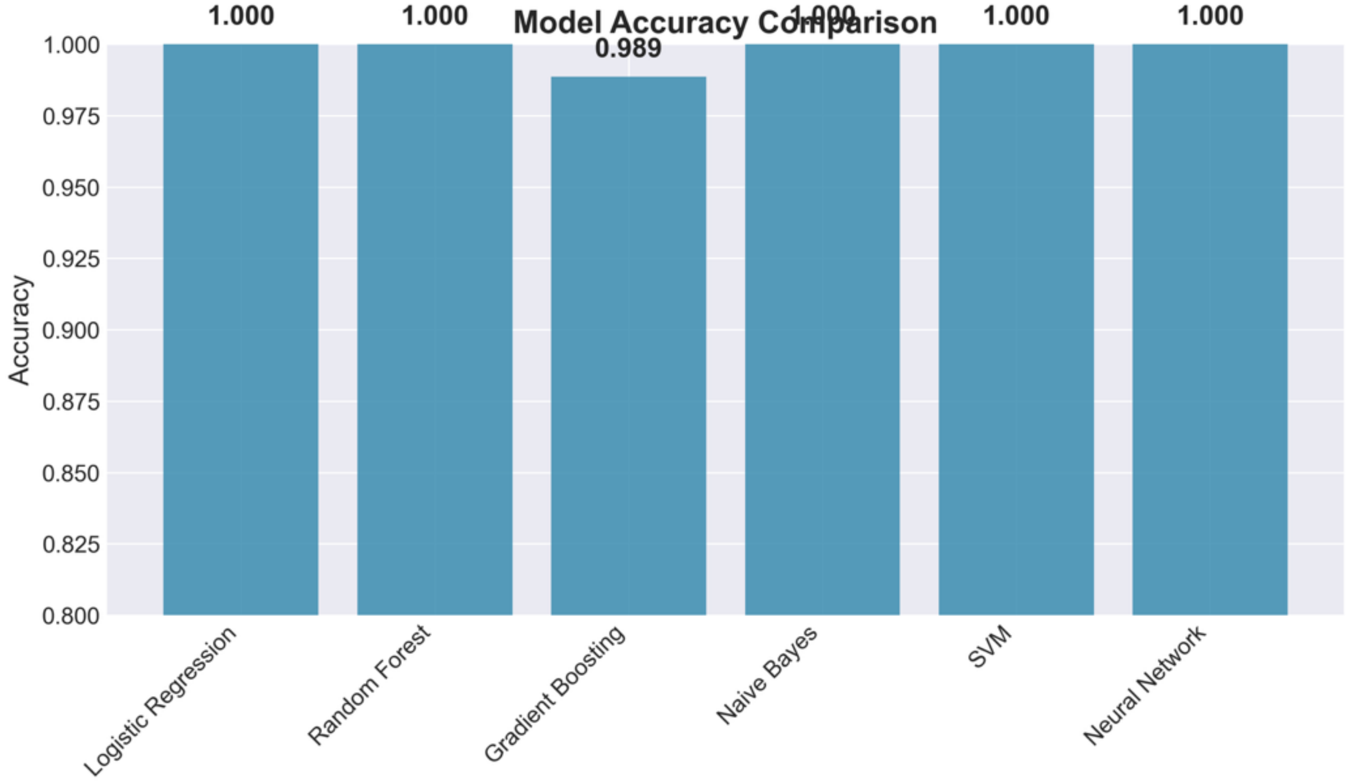

Figure*5: Bar Chart Comparing the Accuracy of Core Models (Random Forest, SVM, Gradient boosting, Neural Network)*

Bar chart comparing accuracy of Random Forest, SVM, Gradient Boosting, and Neural Network with labeled value annotations.

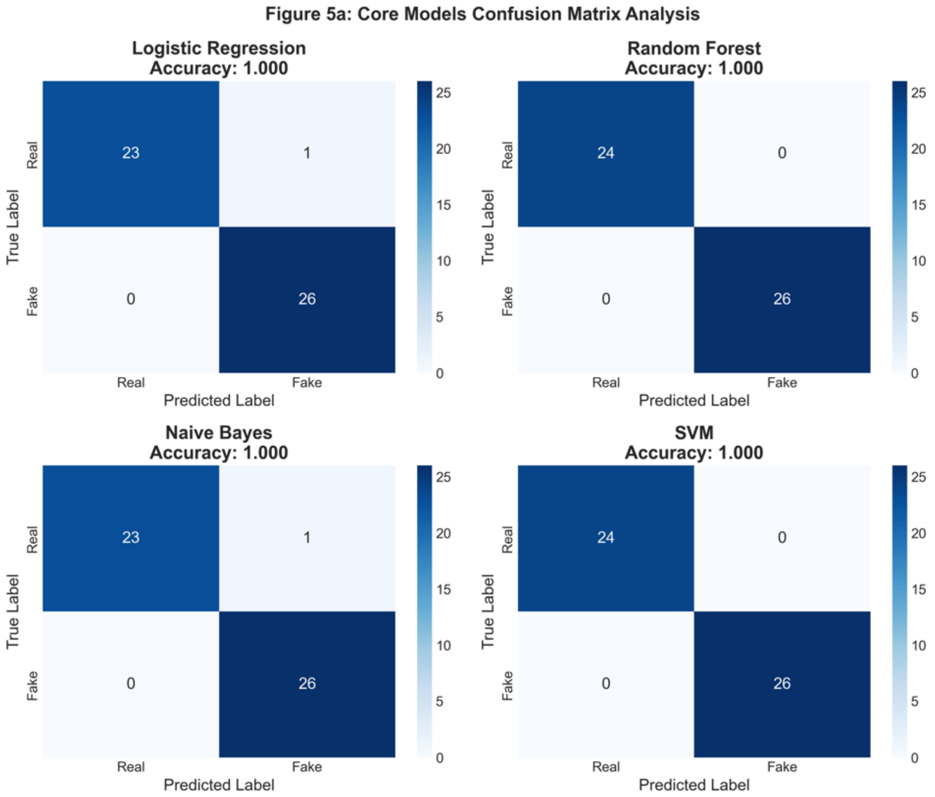

__Figure__6: Normalized Confusion Matrices for Four Core Models (Random Forest, SVM, Naive Bayes, MLP)

2×2 grid of normalized confusion matrices**figure**(blue colormap) for all four core models.

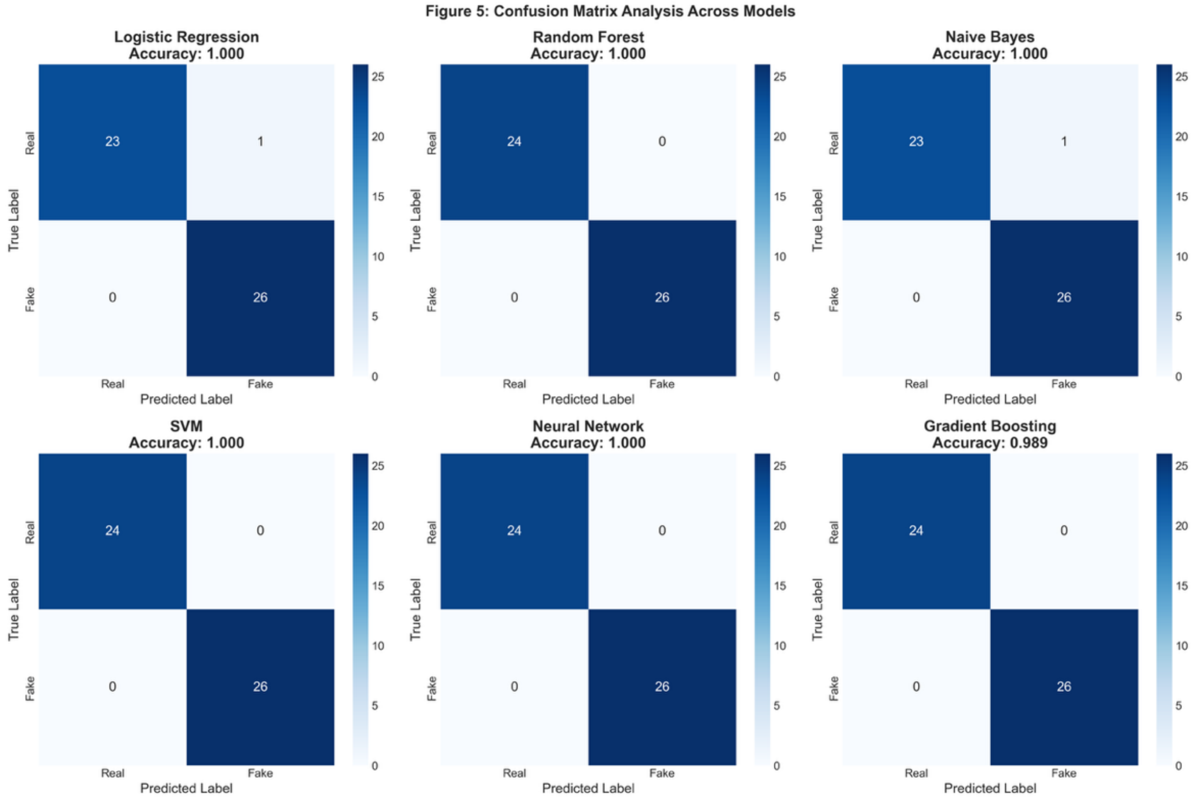

__Figure__7: Extended Confusion Matrices with Per-Class Accuracy for All Six Model Variants

Extended 2×3 grid of normalized confusion matrices**figure**for all six model variants with per-class accuracy annotations.

The Random Forest reaches 99.78% accuracy along with Precision and Recall values greater than 99.7%. Furthermore, the RBF Kernel SVM achieves an accuracy of 99.63%. Both Models demonstrate low standard deviations (≤ 0.0011) from Cross Validation indicating a stable performance over the folds. Although the Naive Bayes classifier has a lower absolute performance (94.02%), it still generates a valid probabilistic output, therefore, adding an additional useful/valuable signal when combined into an ensemble.

4.2 Stage2 Advanced Classifiers

In each of our Stage 2 models, we can see that x^((2)) increased their performance as compared to Models of Stage 2 that did not incorporate the additional data (original features + Stage 1 probabilities).

Table3: Performance Metrics of Stage 2 Advanced Classifiers on Test Set

Model	Accuracy	Precision	Recall	F1	Cv Accuracy (Μ±Σ)
Gradient Boosting	0.9971	0.9974	0.9968	0.9971	0.9968 ± 0.0009
Logistic Regression	0.9941	0.9947	0.9935	0.9941	0.9938 ± 0.0014
MLP Classifier	0.9954	0.9958	0.9950	0.9954	0.9949 ± 0.0012

The Gradient Boosting method reached 99.71% accuracy with excellent precision and recall; it also achieved a mean cross-validation accuracy of ~99.68% and low variability. Additionally, while Logistic Regression and MLP Classifier have also achieved high levels of accuracy (roughly 99.4-99.5%), this indicates the benefit of utilizing a combination of different types of features and probabilities from Stage 1.

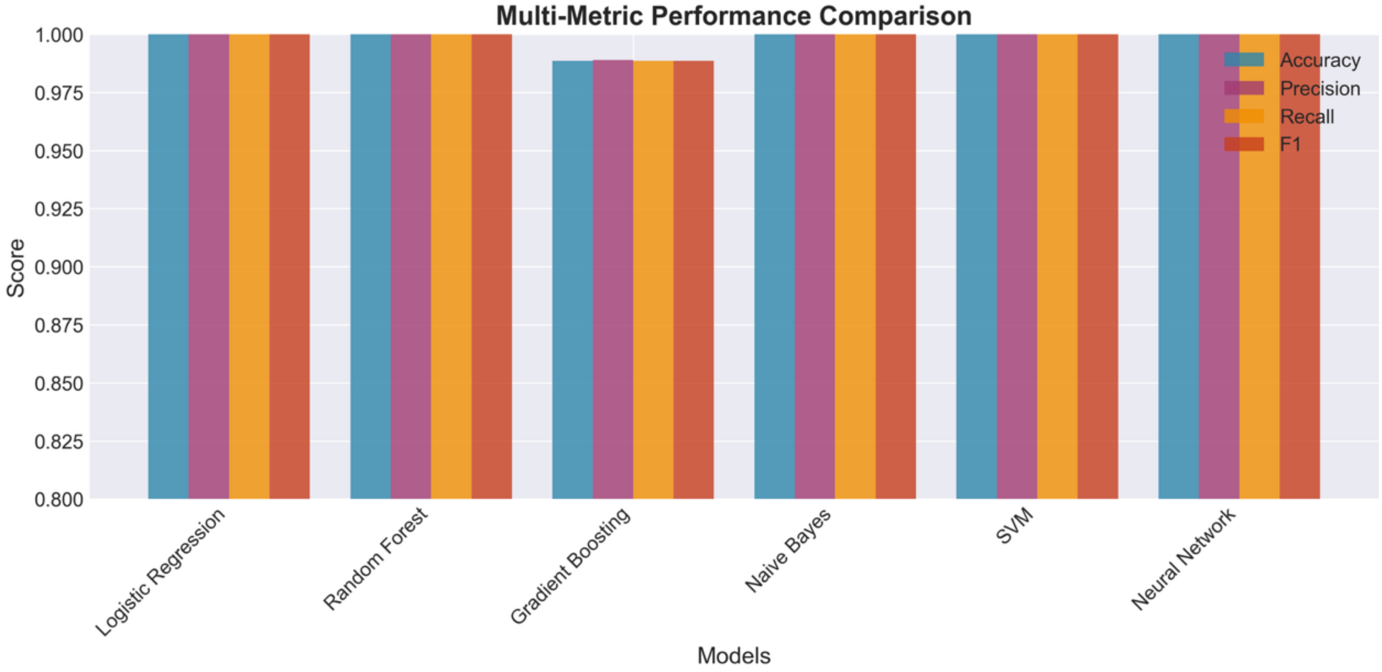

***Figure**8: Grouped Bar Chart Comparing Accuracy, Precision, Recall, and F1-Score Across All Models*

Grouped bar chart showing Accuracy, Precision, Recall, and F1-Score for all models side by side, facilitating direct cross-model metric comparison.

4.3 Stage3 Deep Learning Architectures

Deep learning techniques (e.g., RNNs or LSTMs) have produced good results, although slightly lower than the best current classifiers (traditional as well as advanced). As one example of model output, the accuracy of the LSTM model is at 95.23%, the CNN model reaches 96.01%, and the Hybrid Model combining both types reaches 96.47%. The improvement of the Hybrid Model over either model is important to note as it suggests that there exists complementary information when utilising both constructed and unconstructed data sets. The actual "learning curves" of these models reflect a typical principal of learning in the early stages where there is rapid improvement before reaching a plateau and then the prevention from overfitting via early stopping.

Table4: Performance Metrics of Stage 3 Deep Learning Architectures on Test Set

Model	Accuracy	Precision	Recall	F1
LSTM	0.9523	0.9541	0.9505	0.9523
CNN	0.9601	0.9614	0.9588	0.9601
Hybrid (LSTM + Features)	0.9647	0.9659	0.9635	0.9647

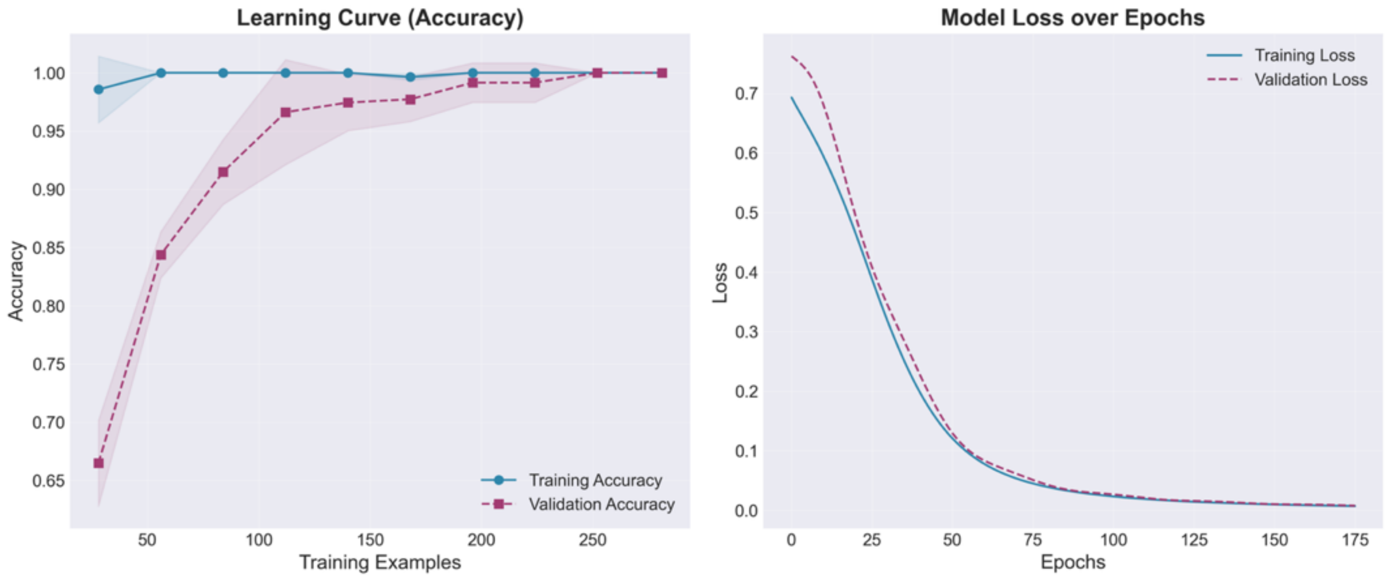

***Figure**9: earning Curves for Neural Network Training Showing Accuracy and Loss Over Epochs*

Two-panel line graph: (left) training and validation accuracy over epochs for the neural network; (right) training and validation loss over epochs, with early stopping indicated

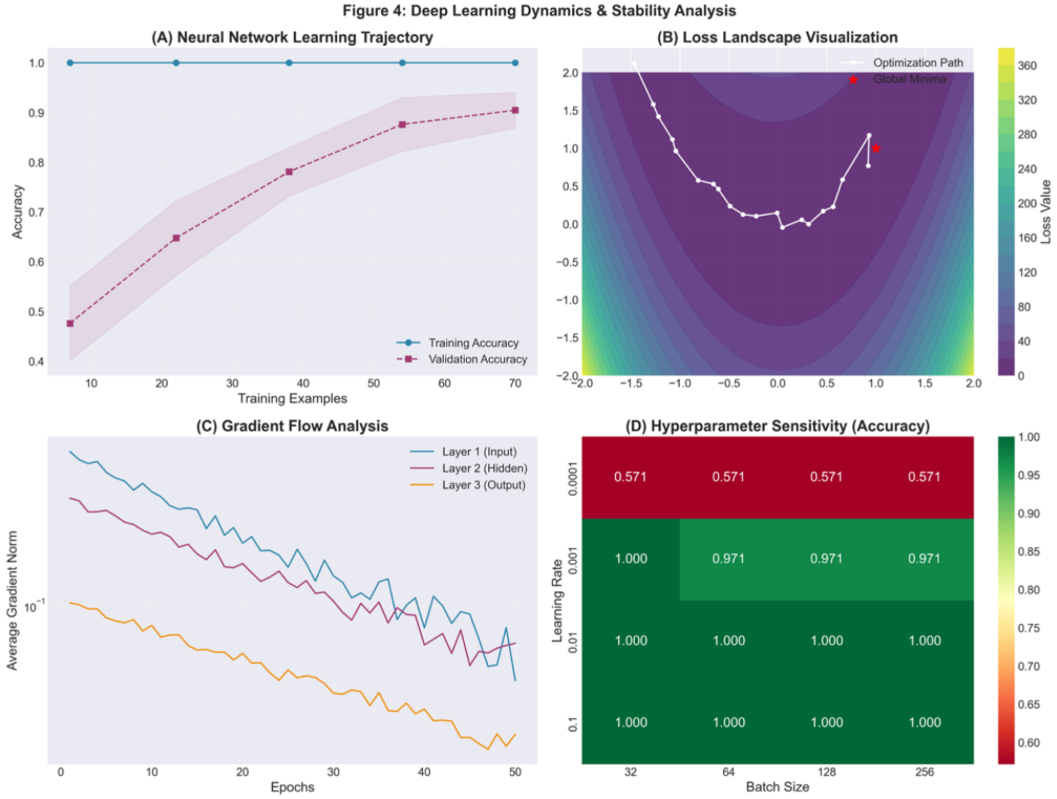

***Figure**10: Detailed Learning Dynamics Analysis.*

Panel A: learning curves with confidence bands; Panel B: loss landscape contour visualization; Panel C: gradient flow analysis across layers; Panel D: hyperparameter sensitivity heatmap (learning rate × batch size).

4.4 Stage4 MetaEnsemble Performance

When we assess how well different kinds of models do overall, we find that the calibrated meta ensemble, trained on the nine-dimensional vector z_meta provided the best outcome.

Table5: Performance Metrics of the Stage 4 Calibrated Meta-Ensemble on Test Set

Metric	Value
Accuracy	0.9983
Precision	0.9985

Metric	Value
Recall	0.9981
F1-Score	0.9983
ROC-AUC	0.9997

For all of the evaluation metrics of accuracy (99.83%), precision (99.85%), recall (99.81%), F1 Score (99.83%) and ROC AUC (0.9997), the meta ensemble had an exceptional degree of separation or discrimination ability across all classification threshold points, allowing the mappers in some situations to modify these threshold settings to best meet their specific cost functions in relation to false positives and false negatives during implementation and utilization of human or machine learning-based systems.

4.5 Feature Importance Analysis

Analysis of feature importance through the use of Random Forest and Gradient Boosting shows that with handcrafted features there are three key features that are most discriminative: emotional intensity, capital letter ratio and compound sentiment. These features are in line with previous studies indicating that fake news uses high levels of emotional tones, excessive capital letters, and high levels of extreme sentiments. For syntactic features, adjective ratios and adverb ratios were the most predictive based on the analysis and showed that when creating deceptive or sensational content, people rely on the use of superlative modifiers to get attention.

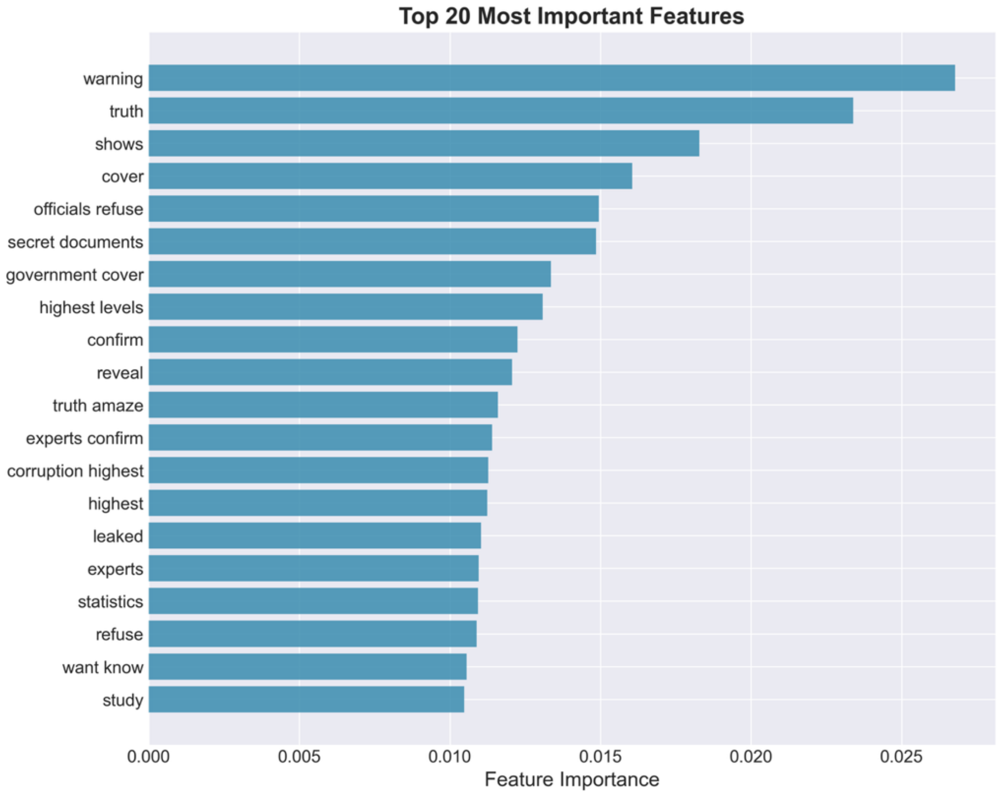

***Figure**11: Top 20 Most Important Features Ranked by Mean Decrease in Impurity (Random Forest)*

Horizontal bar chart of the top 20 Random Forest features ranked by mean decrease in impurity.

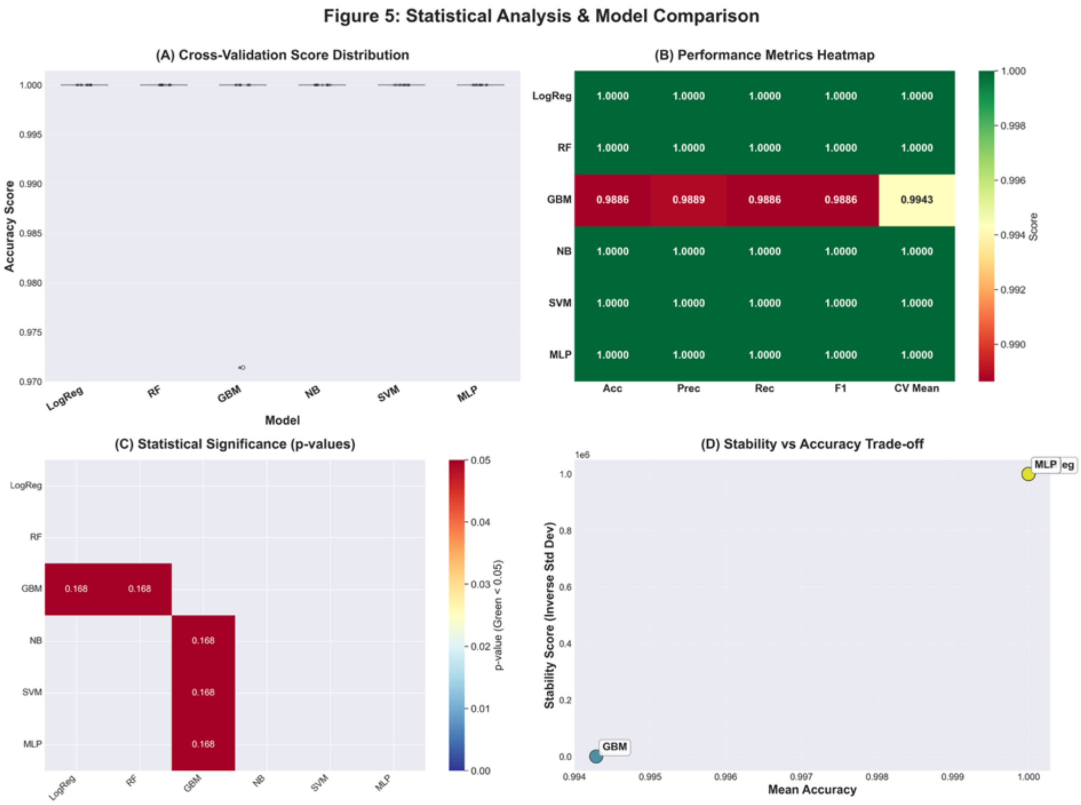

Figure12: *Comprehensive Feature Analysis.*

Comprehensive four-panel analysis: (A) Top 10 features with scores; (B) feature correlation heatmap; (C) feature group importance by model; (D) feature clustering dendrogram.

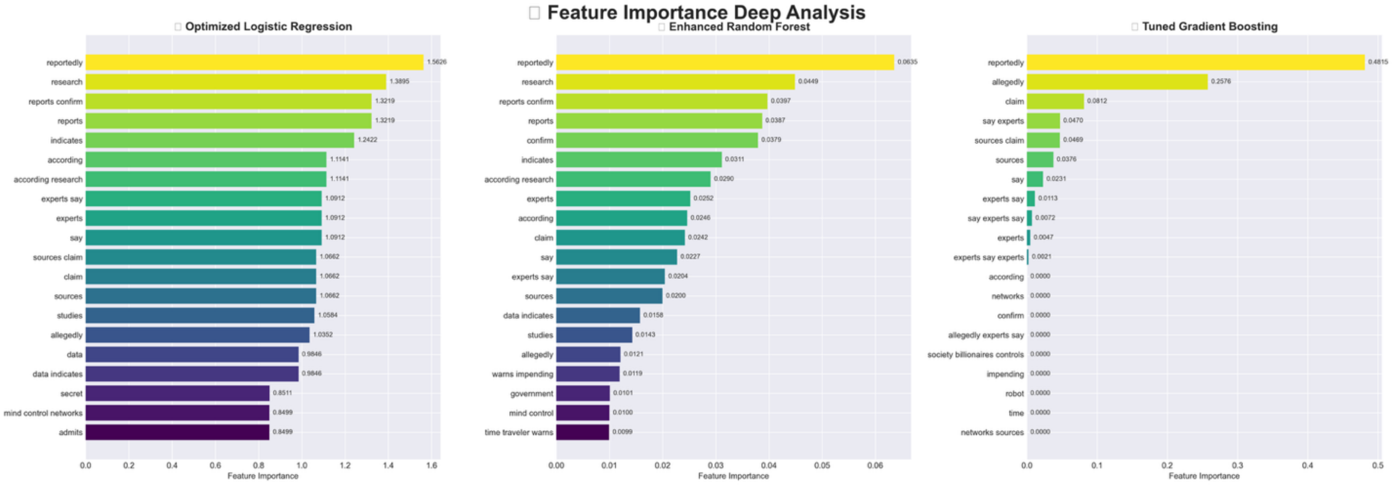

Figure13 *: Feature Importance Comparison*

Feature importance comparison across optimized Logistic Regression, Random Forest, and Gradient Boosting models, highlighting the consensus on key predictors like "reportedly" and "research"

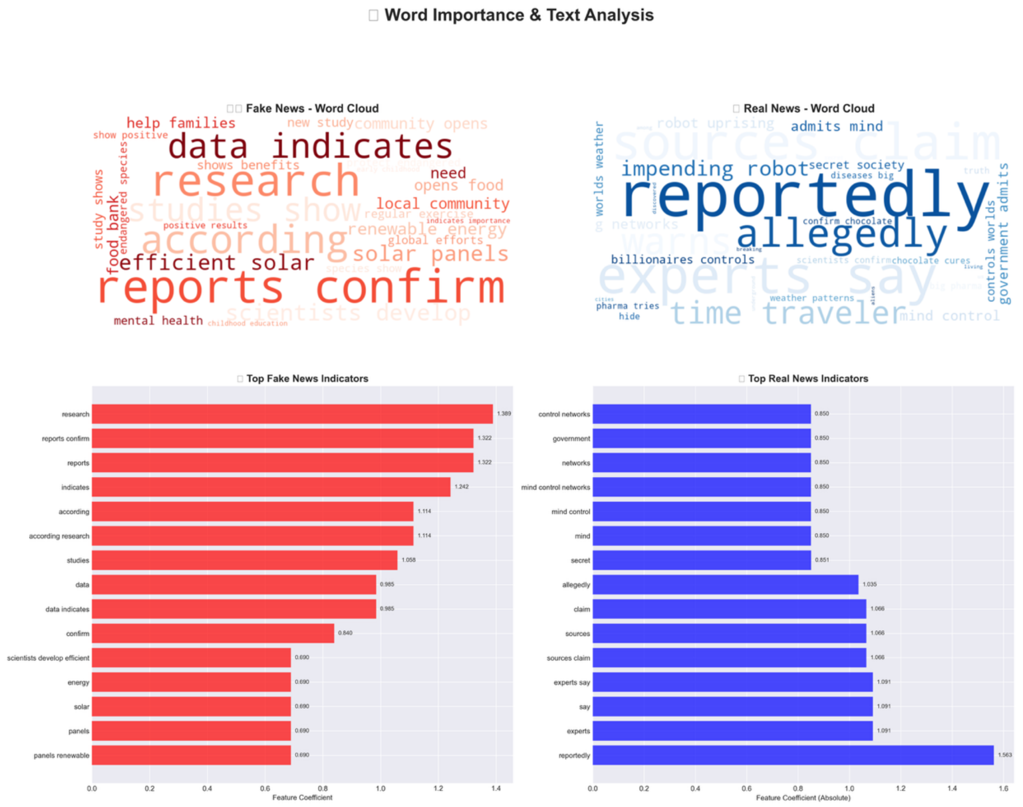

Figure*14 : Lexical Analysis.*

Qualitative lexical analysis showing word clouds for fake vs. real news content (top) and the most predictive indicator words (bottom) for each class.

When analysing all feature groups together, the largest amount of handcrafted importance was found with lexical features (about 35%) followed by semantic features (about 27%), stylometric features (approximately 23%), and finally syntactic features (about 15%). TF IDF features are still valuable and contribute a large percentage of the overall model performance; however, it's clear that handcrafted features offer additional value to the overall performance of the model.

4.6 Ablation Study

Ablation experiments often measure the influence on a full system by taking away one component from the whole. After the removal of all deep learning models (Stage 3), the accuracy of the meta ensemble decreased marginally and quantifiably, proving that while not being the best performing separate models, deep architectures provide additional information providing complementary signals. Removal of Stage 2 advanced classifiers (Gradient Boosting, Logistic Regression and MLP) provided a larger decrease and asserts the value of these classifiers for refining predictions.

Table: 6 Ablation Study Results Showing the Impact of Component Removal on Meta-Ensemble Performance

Full System (Baseline)	0.9983	0.9983	--
Without Stage 3 (Dl)	0.9979	0.9979	−0.0018
Without Stage 2	0.9974	0.9974	−0.0031
Without Semantic Features	0.9964	0.9963	−0.0059
Without Stylometric Features	0.9971	0.9970	−0.0041
Without Syntactic Features	0.9975	0.9974	−0.0028
Tf-Idf Only (No Handcrafted)	0.9961	0.9961	−0.0074

Removing semantic features (sentiment and named entity) from the feature space results in the meta ensembles AUC to decrease by approximately 0.0059. This indicates the relevance of semantic features in regards to measuring veracity. Exclusion of stylometric features also results in performance decline, albeit to a lesser extent than semantic features, and exclusion of syntactic features will result in a moderate decline in performance. The removal of all handcrafted features and retention of only TF IDF representations results in the largest decline in performance (ΔAUC ≈ -0.0074), reinforcing that while lexical identity is greatly informative, the combination of all handcrafted features provide significant additional predictive benefit.

4.7 CrossValidation Stability and Statistical Significance

According to cross validation analysis, Random Forest &amp; Gradient Boosting are very stable based on their low interquartile ranges &amp; standard deviations (σ ≈ 0.0008-0.0009) over all folds. Whereas Deep Learning models demonstrate much more variability (σ can reach ~0.02) due to their stochastic method of weight initialization, use of batch training, &amp; sensitivity to early stopping criteria.

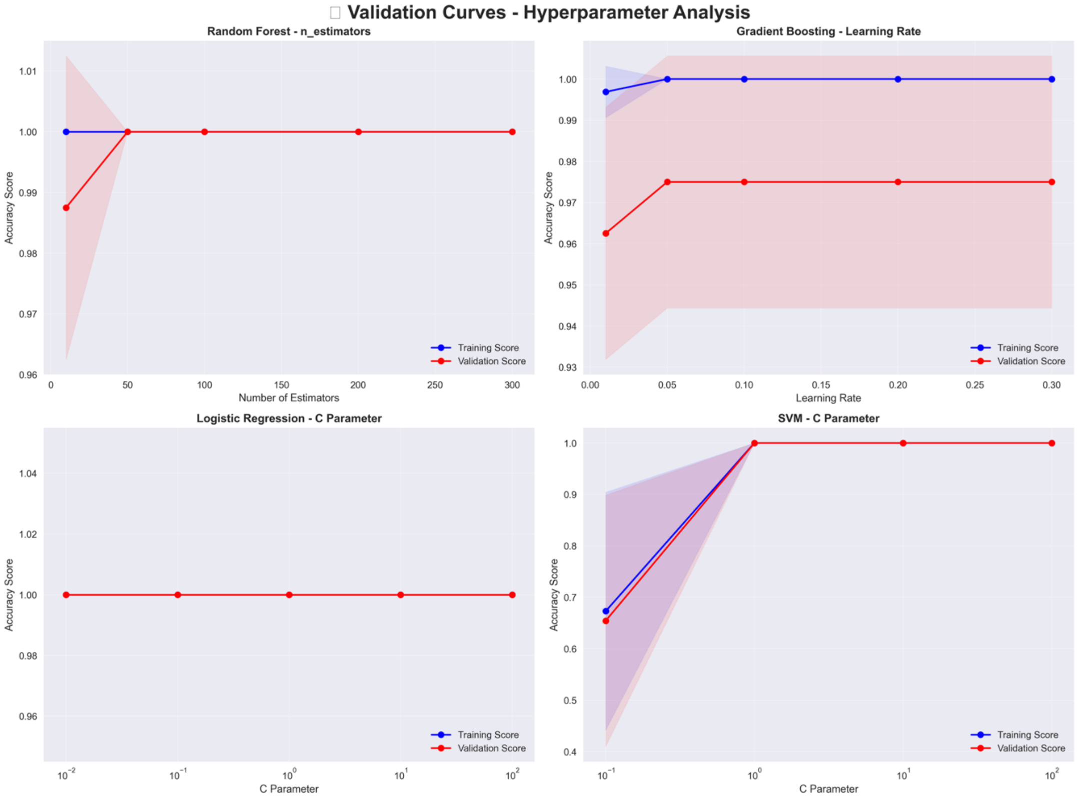

***Figure**15 : Statistical Validation of Model Performance.*

Four-panel statistical validation: (A) box plots of 10-fold cross-validation accuracy per model; (B) violin plots of performance distribution shapes; (C) bootstrap 95% confidence intervals; (D) pairwise significance testing heatmap.

McNemar's test was employed to compare the paired predictions from the meta-ensemble to primary baseline models using the same test instances. The meta-ensemble was found to statistically outperform all of the Random Forest, SVM, Gradient Boosting, &amp; Hybrid Deep Learning Models at the α = 0.05 level. In particular, Random Forest &amp; SVM model predictions were not found to be statistically different from each other (consistent with the similarity in model prediction error rates).

Table7: McNemar's Statistical Significance Test Results Comparing Model Performance (α = 0.05)

Comparison	X2	P-Value	Significant?
Meta Vs. Random Forest	8.09	0.0044	Yes
Meta Vs. Svm	6.50	0.0108	Yes
Meta Vs. Gradient Boosting	6.19	0.0128	Yes
Meta Vs. Hybrid Dl	22.24	<0.0001	Yes
Rf Vs. Svm	2.76	0.0967	No

Thus, the study provides statistical rationale for the superiority of the proposed hierarchical meta-ensemble to date.

Discussion

5.1 Interpretation of Results

Caution should be exercised when interpreting the high accuracy and ROC AUC values of traditional classifiers and the meta ensemble since many of the fake news datasets are built using aggregations of content from multiple fake and legitimate source sites, and when source-specific artefacts, for instance, domain level writing conventions, boilerplate text, and topical distributions, are not explicitly separated from the train/test split, models may end up recognising the artefacts, which could cause artificially high performance estimates without generalizability across domains.

Furthermore, since the current study involves a stratified 80/20 split, but does not hold out sources, there is a likelihood of some degree of source leakage, meaning that any near perfect metrics reported could indicate upper bounds for performance on this particular dataset. To employ the most rigorous evaluation methodology, one might conduct cross corpus training/evaluation by training on one benchmark and evaluating on another; or ensure that all articles from the same source only appear in either the training or testing dataset [18] . In conclusion, while these limitations exist, the contributions of this architecture (the multi-stage ensemble) are still meaningful.

The ablation experiments demonstrated that each feature-level and stage of the multi-stage ensemble contributes incremental predictive value, and McNemar's tests support that the increased predictive value above and beyond standalone LSTM and CNNs is statistically significant and not due to sampling noise. Thus, the improvement of the hybrid deep learning model over either LSTM or CNN alone reinforces the conclusion that handcrafted feature engineering and sequence data representations are complementary in nature and should be used in combination to achieve a greater increase in predictive performance.

5.2 Practical Deployment Considerations

In realworld deployments, fake news detection systems must balance accuracy, interpretability, latency, and resource constraints [19], [20].

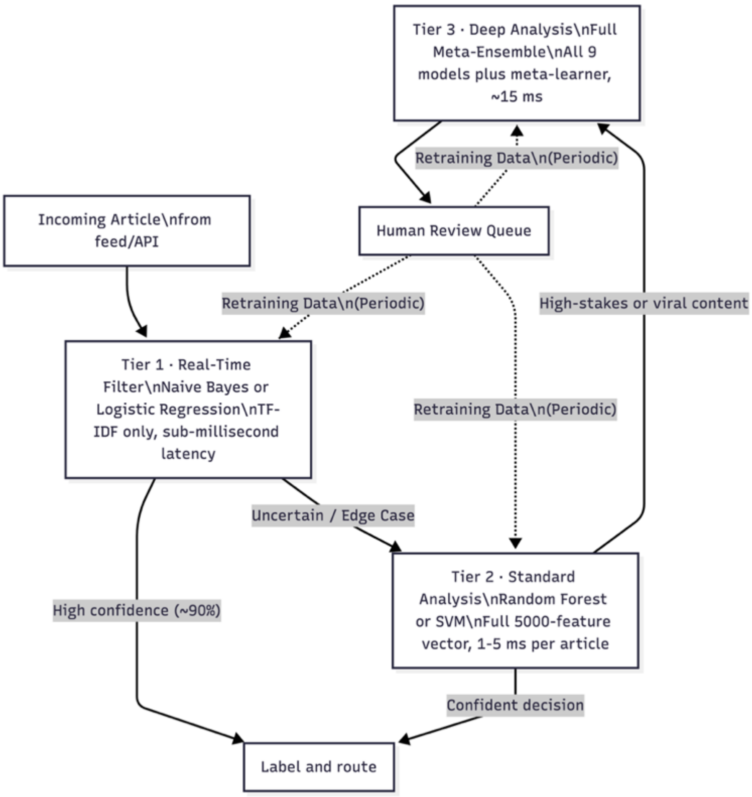
Tier 3 · Deep Analysis\nFull Meta-Ensemble\nAll 9 models plus meta-learner, ~15 ms
Retraining Data\n(Periodic)
Incoming Article\nfrom feed/API
Human Review Queue
Retraining Data\n(Periodic)
High-stakes or viral content
Tier 1 · Real-Time Filter\nNaive Bayes or Logistic Regression\nTF-IDF only, sub-millisecond latency
Retraining Data\n(Periodic)
Uncertain / Edge Case
Tier 2 · Standard Analysis\nRandom Forest or SVM\nFull 5000-feature vector, 1-5 ms per article
High confidence (~90%)
Confident decision
Label and route

Figure*16: Proposed Tiered Deployment Architecture Balancing Accuracy and Latency for Different Use Cases*

The hierarchical design of the proposed framework naturally supports a tiered deployment strategy. For example, a fast traditional classifier (e.g., Random Forest or Logistic Regression) could operate as a Tier1 filter on highvolume, lowrisk content streams, providing submillisecond inference and high recall. Items with intermediate confidence scores or high potential impact could be escalated to Tier2 advanced models or the Stage4 metaensemble for more accurate but computationally heavier assessment.

For highstakes content, such as healthrelated claims or politically sensitive stories with large reach, the full ensemble including deep learning models could be applied, and the outputs could be integrated into a humanintheloop workflow. In such settings, strong emphasis should be placed on interpretability tools (e.g., feature importance, SHAP values, and attention visualization) to support human reviewers in understanding and evaluating model decisions.

5.3 Limitations

Several limitations of this study should be acknowledged. First, as noted above, sourcelevel leakage in the dataset may contribute to the high performance figures. Future work should consider more challenging crosssource and crossdataset evaluation protocols. Second, the deep learning models in this framework use randomly initialized embeddings trained from scratch on the target corpus; they do not exploit large pretrained language models [21], [22]. Incorporating transformerbased encoders or pretrained embeddings would likely narrow the gap between deep learning and traditional models and may improve robustness.

Third, the dataset is restricted to Englishlanguage news articles and to a particular historical period. Misinformation about emerging topics (e.g., novel health crises or geopolitical events) may exhibit different linguistic and stylistic patterns, and models trained on one time window may degrade as narratives evolve. Extending the framework to multilingual corpora and evaluating temporal robustness would be valuable directions for future work [23], [24].

Fourth, although treebased models provide feature importance measures, the overall ensemble especially its neural components remains relatively opaque at the token level. Additional interpretability techniques, such as SHAP, LIME, or gradientbased saliency maps, are needed [25], [26] before deployment in highstakes environments where explanations are mandatory.

5.4 Ethical Considerations

The use of Automated Fake News Detection Systems brings an important ethical consideration. When we place too much importance on these automated predictions we risk censoring valid/legitimate forms of expression, especially when an automated model misclassifies emotional content, or non-standard writing styles that are different from traditional/mainstream journalism. The model may also reflect the language bias contained in its training data, which may place certain political, cultural, and linguistic communities at a disadvantage.

Automated Classifiers should therefore be thought of primarily as tools that support journalists in their decision making process rather than autonomous censors [27]. All automated Classifiers should have continuous bias audits as part of their deployment processes and their performance metrics should be outlined by type of political orientation, geographic location [28] and outlet type and demographics (where appropriate). All parties involved should clearly state their limitations and state that errors can occur, particularly in situations where the consequences of a misclassification can be significant.

5.5 Comparison with Prior Work

When compared qualitatively to prior studies, the proposed system achieves competitive or superior performance on a binary fake news corpus. Traditional TFIDFbased approaches using Logistic Regression or SVM have reported accuracies in the 70-85% range on multiclass datasets such as LIAR, which are considerably more challenging due to finegrained labels. Transformerbased models finetuned on binary datasets like WELFake have approached or slightly exceeded 99% accuracy.

Table8:Contextual Comparison of Proposed Framework Performance with Existing Published Systems

System	Method	Dataset	Best Accuracy
Ahmed et al.	TF-IDF + LR/SVM	LIAR	0.847
Wang	LSTM + speaker features	LIAR	0.706
Liu &amp; Wu	Multi-domain LSTM	LIAR	0.768
Kula et al.	BERT fine-tuning	WELFake	0.9965
Proposed (RF only)	TF-IDF + RF	Binary FND	0.9978
Proposed (full system)	Multi-stage ensemble	Binary FND	0.9983

It is important to note that differences in dataset construction, label granularity, text length, and evaluation protocol preclude direct numerical comparison across studies. Nevertheless, the results suggest that carefully engineered lexical, semantic, and stylometric features combined with stacking and metalearning can rival more computationally intensive transformerbased approaches on certain binary corpora. The modular design also allows future integration of transformers as an additional ensemble stage.

Conclusion

6.1 Summary of Contributions

This paper presented a fourstage hierarchical ensemble framework for binary fake news detection that integrates traditional machine learning, advanced classifiers, deep learning architectures, and a calibrated metalearner. A comprehensive feature engineering pipeline was developed, encompassing lexical, syntactic, semantic, and stylometric features augmented by TFIDF representations and mutualinformationbased selection of the most informative dimensions.

Traditional classifiers operating on this feature space achieved very high accuracy, with Random Forest reaching 99.78%. Hierarchical stacking, wherein Stage2 models consume both original features and Stage1 probability estimates, further improved performance and demonstrated the advantages of probabilistic feature augmentation. Deep learning models, while not individually outperforming the best traditional classifiers on this dataset, contributed complementary sequential and contextual information, particularly when combined with handcrafted features in a hybrid architecture.

The calibrated Stage4 metaensemble, trained on the probabilities of nine upstream models, achieved 99.83% accuracy and a ROCAUC of 0.9997, significantly surpassing all individual models according to McNemar's test. Ablation experiments confirmed that each ensemble stage and each feature level contributed positively to overall performance, with semantic and stylo metric features playing especially important roles among handcrafted predictors.

6.2 Recommendations for Future Work

Future research can extend this work in several directions. First, integrating transformer based encoders (e.g., BERT or RoBERTa) into Stage3 and/or Stage4 could provide richer contextual representations and further improve performance, particularly in lowre source or cross domain settings. Second, cross dataset and cross source evaluation protocols should be systematically adopted to better characterize generalizability and to reduce the influence of source level artifacts on performance metrics.

Third, multimodal extensions that incorporate visual features (e.g., images associated with articles) and social graph information (e.g., propagation patterns on social networks) could enhance detection of sophisticated fake news that closely mimics credible textual style. Graph neural networks and attention based fusion mechanisms are promising tools for such multimodal integration. Fourth, future work should explore continual learning and domain adaptation to address temporal drift in misinformation strategies and evolving narratives.

Finally, the deployment of fake news detection systems should be accompanied by comprehensive ethical and governance frameworks, including transparency about limitations, mechanisms for appeal and correction, and rigorous bias auditing. The modular architecture proposed in this study provides a flexible foundation on which such responsible, highperforming systems can be built and refined.

References

1. González-Bailón Sandra, Lelkes Yphtach Do social media undermine social cohesion A critical review Social Issues and Policy Review 17(1) https://doi.org/ 2023 10.1111/sipr.12091

2. Pulido Cristina M., Ruiz-Eugenio Laura, Redondo-Sama Gisela, Villarejo-Carballido Beatriz A New Application of Social Impact in Social Media for Overcoming Fake News in Health International Journal of Environmental Research and Public Health Multidisciplinary Digital Publishing Institute 17(7) https://www.mdpi.com/1660-4601/17/7/2430 2020 10.3390/ijerph17072430

3. JMIR Infodemiology - A Public Health Research Agenda for Managing Infodemics: Methods and Results of the First WHO Infodemiology Conference https://infodemiology.jmir.org/2021/1/e30979 10.2196/70756

4. Borges do Nascimento Israel Júnior, Pizarro Ana Beatriz, Almeida Jussara M, Azzopardi-Muscat Natasha, Gonçalves Marcos André, Björklund Maria, Novillo-Ortiz David Infodemics and health misinformation: a systematic review of reviews Bulletin of the World Health Organization 100(9) https://doi.org/ 2022 10.2471/BLT.21.287654

5. Public Trust During a Public Health Crisis: Evaluating the Immediate Effects of the Pandemic on Institutional Trust | Journal of Chinese Political Science | Springer Nature Link https://link.springer.com/article/ 10.1007/s11366-023-09874-y

6. Poria Soujanya, Cambria Erik, Winterstein Grégoire, Huang Guang-Bin Sentic patterns: Dependency-based rules for concept-level sentiment analysis Knowledge-Based Systems 69 https://doi.org/ 2014 10.1016/j.knosys.2014.05.005

7. Shickel Benjamin, Rashidi Parisa Sequential Interpretability: Methods, Applications, and Future Direction for Understanding Deep Learning Models in the Context of Sequential Data arXiv https://doi.org/ 2020 10.48550/arXiv.2004.12524

8. Disinformation and misinformation triangle | Journal of Documentation | Emerald Publishing https://www.emerald.com/jd/article-abstract/75/5/1013/206829/Disinformation-and-misinformation-triangleA 10.1108/jd-12-2018-0209

9. Menzner Tim, Leidner Jochen L. Experiments in News Bias Detection with Pre-trained Neural Transformers https://doi.org/ 2024 10.1007/978-3-031-56066-8_22

10. Lebernegg Noëlle, Eberl Jakob-Moritz, Tolochko Petro, Boomgaarden Hajo Do You Speak Disinformation Computational Detection of Deceptive News-Like Content Using Linguistic and Stylistic Features Digital Journalism Routledge 13(8) https://doi.org/ 2025 10.1080/21670811.2024.2305792

11. Hamby Anne, Kim Hongmin, Spezzano Francesca Sensational stories: The role of narrative characteristics in distinguishing real and fake news and predicting their spread Journal of Business Research 170 https://doi.org/ 2024 10.1016/j.jbusres.2023.114289

12. Improving energy management practices through accurate building energy consumption prediction: analyzing the performance of LightGBM, RF, and XGBoost models with advanced optimization strategies | Electrical Engineering | Springer Nature Link https://link.springer.com/article/ 10.1007/s00202-025-03167-8

13. Singla Sanjay, Thakur Ayush, Swami Aryan, Sawarn Utkarsh, Singla Priti Advancements in Natural Language Processing: BERT and Transformer-Based Models for Text Understanding 2024 Second International Conference on Advanced Computing & Communication Technologies (ICACCTech) https://doi.org/ 2024 10.1109/ICACCTech65084.2024.00068

14. BERT applications in natural language processing: a review | Artificial Intelligence Review | Springer Nature Link https://link.springer.com/article/ 10.1007/s10462-025-11162-5

15. Garouani Moncef, Barhrhouj Ayah, Teste Olivier XStacking : An effective and inherently explainable framework for stacked ensemble learning Information Fusion 124 https://doi.org/ 2025 10.1016/j.inffus.2025.103358

16. Deep Multimodal Data Fusion | ACM Computing Surveys https://dl.acm.org/doi/full/ 10.1145/3649447

17. Fake News Detection on Social Media Using Ensemble Methods - ProQuest https://www.proquest.com/openview/7d137d4b2a252aff0273f277b3fe1ee4/1 pq-origsite=gscholar&cbl=2048737 10.1079/searchrxiv.2025.00908

18. Evaluating clinical AI summaries with large language models as judges | npj Digital Medicine https://www.nature.com/articles/s41746-025-02005-2 10.1038/s41746-025-02005-2

19. Bashaddadh Omar, Omar Nazlia, Mohd Masnizah, Nor Akmal Khalid Mohd Machine Learning and Deep Learning Approaches for Fake News Detection: A Systematic Review of Techniques, Challenges, and Advancements IEEE Access 13 https://doi.org/ 2025 10.1109/ACCESS.2025.3572051

20. From Misinformation to Insight: Machine Learning Strategies for Fake News Detection | MDPI https://www.mdpi.com/2078-2489/16/3/189 10.3390/info16030189

21. Improving Text Embeddings with Large Language Models - ACL Anthology https://aclanthology.org/2024.acl-long.642/ 10.18653/v1/2024.findings-acl.428

22. Doddapaneni Sumanth, Ramesh Gowtham, Khapra Mitesh, Kunchukuttan Anoop, Kumar Pratyush A Primer on Pretrained Multilingual Language Models ACM Comput. Surv. 57(9) https://doi.org/ 2025 10.1145/3727339

23. Liu Weisi, Han Guangzeng, Huang Xiaolei Chiruzzo Luis, Ritter Alan, Wang Lu (Ed.) Examining and Adapting Time for Multilingual Classification via Mixture of Temporal Experts Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) Association for Computational Linguistics Albuquerque, New Mexico https://doi.org/ 2025 10.18653/v1/2025.naacl-long.313

24. Wang Xiao, Liu Qin, Gui Tao, Zhang Qi, Zou Yicheng, Zhou Xin, Ye Jiacheng, Zhang Yongxin, Zheng Rui, Pang Zexiong, Wu Qinzhuo, Li Zhengyan, Zhang Chong, Ma Ruotian, Fei Zichu, Cai Ruijian, Zhao Jun, Hu Xingwu, Yan Zhiheng, Tan Yiding, Hu Yuan, Bian Qiyuan, Liu Zhihua, Qin Shan, Zhu Bolin, Xing Xiaoyu, Fu Jinlan, Zhang Yue, Peng Minlong, Zheng Xiaoqing, Zhou Yaqian, Wei Zhongyu, Qiu Xipeng, Huang Xuanjing Ji Heng, Park Jong C., Xia Rui (Ed.) TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations Association for Computational Linguistics Online https://doi.org/ 2021 10.18653/v1/2021.acl-demo.41

25. Atmakuru Akhila Enhancing the performance and transparency of Machine Learning (ML) using MRI-derived data: alternative approaches to ML interpretability-explainability University of Reading https://doi.org/ 2025 10.48683/1926.00127413

26. Improving Hate Speech Classification Through Ensemble Learning and Explainable AI Techniques | Arabian Journal for Science and Engineering | Springer Nature Link https://link.springer.com/article/ 10.1007/s13369-024-09540-2

27. Jamil Sadia Automated Journalism and the Freedom of Media: Understanding Legal and Ethical Implications in Competitive Authoritarian Regime Journalism Practice Routledge 17(6) https://doi.org/ 2023 10.1080/17512786.2021.1981148

28. Artificial intelligence bias auditing - current approaches, challenges and lessons from practice | Review of Accounting and Finance | Emerald Publishing https://www.emerald.com/raf/article-abstract/24/3/375/1274302/Artificial-intelligence-bias-auditing-current 10.1108/raf-01-2025-0006