iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74154

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Linguistic Feature-Driven Fake News Detection

Vineet[1], Sunil Kumar Nandal[2]
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology

Abstract: ***This study presents a data mining-driven prediction approach for assessing the well-being of IT workers. The strategy puts individuals into groups depending on how likely they are to become sick, such low, medium, and high. It looks at factors like how active they are, how they eat, how stressed they are, how well they sleep, and their medical history. The strategy combines appropriate preprocessing, feature selection, and classification algorithms to make sure the findings are correct and trustworthy. There are a number of tables and figures that support the testing findings and show that the model may provide important health information. The findings illustrate how it may help individuals remain healthy at work, recognize threats early, and stay healthy. This approach is meant to bring IT-based decision-making and keeping an eye on workers' health closer together.***

Keywords: Predictive Model, Data Mining, Health Analysis, Risk Classification, Preventive Healthcare, Data Preprocessing, ML, Occupational Health, Wellness Monitoring.

I. INTRODUCTION

In the digital era, social media and internet news sites have made it much easier for information to disseminate quickly. It is now easier for more individuals to manufacture things, which has led to more misleading or wrong information being passed off as news. This has given individuals greater power, but it has also made it easier for false information to propagate. Finding fake news quickly is a significant problem in AI and computational linguistics since it may change people's minds, cause social unrest, and damage democracy. individuals that check for false news in the past have usually looked at things like the content, how individuals respond, and how information spreads via networks. But these plans won't function if fresh stories come out or if facts like network connections and metadata aren't accessible. Things like style, vocabulary, grammar, and semantics may be used instead in this scenario. There are little variations between true and false news that might help you tell them apart. For example, the way the writer utilizes rhetoric, the emotional tone, and how divided the emotions are.

TABLE I
LINGUISTIC FEATURE-DRIVEN FAKE NEWS DETECTION

Aspect	Details
Core Approach	Uses linguistic features such as word frequency, part-of-speech (POS) distribution, readability scores, and discourse structures.
Working Principle	Extracts numerical values from text using NLP techniques and identifies stylistic or structural patterns linked with dishonest writing.
Key Features	- Word frequency analysis - POS distribution - Readability ratings - Discourse structures
Advantage over Fact-checking	Works in real-time without relying on external databases or sources.
Integration with AI	Machine learning (ML) and deep learning (DL) models enhance detection using linguistic data.
Accuracy & Interpretability	Provides high detection accuracy and improves explainability, showing why a story was labeled as fake.
Scalability	Works with any language and is independent of context or situation.
Relevance	Essential as false information spreads rapidly and is harder to detect.
Research Objective	To examine, extract, and use linguistic traits to create a dependable methodology for fake news detection, especially in low-data and dynamic environments.

A. Background

Digital media platforms are developing so swiftly that the methods people create, exchange, and consume information have altered. This modification has made it feasible to communicate to individuals right away and see things from other points of view. But technology has also made it simpler for lies to spread. "Fake news" is news that has been made up or is inaccurate and is then passed off as actual news.

TABLE II
ROLE OF LINGUISTIC FEATURE ANALYSIS IN FAKE NEWS DETECTION

Aspect	Details
Reasons for Fake News Creation	- To sway public opinion - Influence election outcomes - Gain financial benefits
Limitations of Traditional Fact-checking	- Accurate but time-consuming - Cannot keep up with the huge volume of online information
Emergence of Linguistic Feature Analysis	Uses syntax, semantics, discourse patterns, and stylistic indicators to automatically detect fake news.
Supporting Technologies	- Natural Language Processing (NLP) - Machine Learning (ML)
Detection Capability	Finds subtle signals of deception and false information in text.
Significance	Plays a crucial role in combating misinformation circulating online.

B. Motivation of Research

Our work is motivated by a need to address the growing menace of deception in digital communication. This necessity has been the fundamental reason why we did this study. On the other side, fake news has a lot of adverse repercussions on society. For instance, it makes individuals more politically divided, makes them less inclined to believe actual news, and puts their health at danger in crises. There are a number of techniques to tell whether a letter is genuine or fake, but most of them don't work because they don't check for the linguistic fingerprints that make fraudulent letters different from real ones. Language uses psychological, cultural, and cognitive parts to show what a statement means and how real it is. This project seeks to provide a comprehensive and scalable methodology for identifying fake news across several domains and languages by systematically extracting and analyzing linguistic indicators, such as sentiment polarity, lexical variety, part-of-speech patterns, and discourse coherence. The quest to attain this aim is motivated by the desire to discover solutions that are both technically sound and beneficial to society.

C. Contribution of Research

This research makes several key contributions to the field of automated fake news detection:

TABLE III
CONTRIBUTION OF RESEARCH

S. No.	Contribution Area	Description
1	Novel Linguistic Feature Extraction	Developed an advanced framework to extract syntactic, semantic, and stylistic features from news text for accurate fake news identification.
2	Multi-Layer Feature Analysis	Combined multiple linguistic levels (morphological, lexical, and discourse) to enhance detection performance and reduce false positives.
3	Dataset Enrichment	Created or curated domain-specific datasets with annotated fake and real news samples, including linguistic labels for training models.
4	Hybrid Detection Model	Integrated linguistic features with machine learning and deep learning techniques to achieve robust classification results.
5	Cross-Domain Evaluation	Validated the model on multiple domains (politics, health, entertainment) to ensure generalizability of the proposed approach.
6	Real-Time Detection Capability	Designed a scalable, low-latency detection pipeline suitable for integration into online news monitoring systems.
7	Interpretability & Explainability	Enhanced transparency by identifying which linguistic features most influence classification decisions, aiding trust and adoption.
8	Performance Benchmarking	Compared proposed approach with state-of-the-art methods, demonstrating superior accuracy, precision, and recall.

By emphasizing linguistic indicators, this research bridges the gap between computational detection and human-like judgment, paving the way for more interpretable and trustworthy fake news detection systems.

II. LITERATURE REVIEW

Bellam (2025) describe a hybrid pipeline that employs TF-IDF textual representations, bidirectional GRUs, and standard count vectorizers to detect bogus news. Their approach integrates deep sequential modeling (Bi-GRU) with basic lexical characteristics to ascertain word relevance and contextual semantics at the document level. Tests indicated that it functioned better than baselines with just one representation. The pros are that the ablation studies are clear and the feature fusion is realistic. The drawbacks are that language and generalization weren't studied in depth, and the dataset may have some bias. The study shows that we need to look at multimodal extensions and better contextual embeddings [1].

La, et al. (2025) created KGAlign as a means to discover fake news by matching knowledge-graph signals with text and images in different ways. They exhibit good results on multimodal benchmarks and put cross-modal interactions into one representation to discover tiny discrepancies between modalities and outside information. The paper is new since it puts structured knowledge into multimodal pipelines, even if it doesn't work well with big real-world graphs and demands high-quality KGs.[2].

Liu et al. (2025) provide a comprehensive examination of various machine learning algorithms for identifying fraudulent social media activity, including a wide range of methodologies, problems, biases, and classification strategies. The study meticulously examines dataset imbalances, annotation subjectivity, and demographic or platform biases that undermine model performance, while also highlighting evaluation procedures and using supervised, unsupervised, and hybrid techniques. The paper makes clear methodological problems with fairness, interpretability, and robustness against hostile manipulation, while also giving a good picture of the whole picture. It is recommended that defined benchmarks and techniques be implemented to mitigate prejudice [3].

Zhang (2025) introduce LinguaSynth as a framework for categorising news articles, integrating pragmatic, syntactic, and semantic components via varied linguistic data. The method combines different types of linguistic clues into one classifier, which makes it easier to detect news stories that are misleading or ambiguous than just text. Some of the problems are that you need to make sure the results work for multiple languages and that you might be overfitting to signals that are only present in one language. You have to do a lot of feature engineering and validation across datasets, which is one of the perks. The authors say that LinguaSynth works better when it is utilised with contextual embeddings [4].

Choudhary (2024) introduce a graph isomorphic network-based false-news detector named GIN-FND, which employs user preference structures to enhance the classification of social platforms. The technology use isomorphism-aware graph networks to depict interaction graphs and user behaviour patterns, facilitating the distinction between benign sharers and those disseminating misleading information. It does this by measuring propagation and preference homophily. The method works well with social-structural indicators, but it might not perform well if user data privacy rules are strict or if interaction graphs are sparse or noisy. Temporal dynamics and privacy-preserving graph learning are two areas that could be explored in the future [5].

Zhao, et al. (2024) suggest a dual-channel graph convolutional attention network to find fake news. This network looks at content and propagation graphs individually and then integrate the results using attention processes. The model outperforms single-channel baselines due to its dual channels, which enable it to focus on significant nodes and edges across both semantic and social propagation domains. The attention fusion in the design is based on good theory, however I'm worried that the model will be hard to utilise with a lot of data and hard to grasp. They need to do additional research on how to make it work better and how to make it less likely to be re-wired aggressively [6].

Alian et al. (2024) examine feature-driven end-to-end test generation, which is the process of automatically making test cases using extracted features and behavioural signals. Their feature-centric test development methods might not work right away for finding fake news, but they are useful for developing strong evaluation pipelines and testing classifiers with edge-case inputs. The study has some excellent points, such showing how automated pipelines and increased test coverage might assist. But it also has some problematic points, including needing to do feature engineering for each domain and the fact that there might not be any genuine adversaries. Adding these kinds of test frameworks during model validation could make models that find false information more reliable [7].

Wang, et al. (2024) tackle the difficulty of identifying small objects in urban traffic using UAV vision by implementing attention mechanisms and multi-scale feature-driven networks. Their methods for capturing multi-scale context and focused attention can help you identify misleading information in pictures, such when you notice doctored photos or minor visual artefacts. The study shows that UAV datasets improve accuracy, but it also shows that there are trade-offs between processing costs and scale sensitivity. The multi-scale + attention motif should enable multimodal systems that look at pictures on social media to find fake news [8].

Chen (2023) presents an initial observation to determine if a particular linguistic attribute may definitively enhance the precision of false news detection. Certain variables, such as hedging and subjectivity indicators, provide predictive usefulness within particular datasets however do not generalise across domains; this represents one of the advantages and disadvantages emphasised by the study. The contribution promotes ensemble or multi-feature methodologies while warning against over-reliance on single-feature heuristics. Future investigations must validate the results across other languages and platforms, considering the limitations of the datasets and the exploratory essence of the work [9].

Balshetwar, Rs & R (2023) demonstrate that sentiment cues, with textual data, can assist in identifying misinformation in their research on detecting fake news on social media through sentiment-analysis-informed classifier algorithms. They demonstrate through a comparison of various classical classifiers that sentiment-aware features are effective in identifying deceptive information associated with emotional tone, particularly in political or sensationalistic content. However, sentiment alone is inadequate and may be masked by caustic or confusing language; the management of linguistic sentiment variations and the resilience to such nuances remain unresolved challenges [10] .

Yadav et al. (2023) put forward a hybrid deep learning approach that uses the complementary features of different brain modules to find false information. These modules can use CNNs for local semantics and RNNs or transformers for sequence-level context. Their hybrid architecture outperforms single-model baselines on benchmark datasets, and ablation investigations elucidate the contribution of each component. Some of the concerns are how complicated the models are, how much it costs to train them, and maybe how sensitive they are to the amount of the dataset. It can still be better used in situations with fewer resources [11].

Madden (2023) proposes a style-based methodology for identifying COVID-19-related fake news, emphasising stylometry, linguistic style features, and their interplay with epidemic-specific material. This thesis asserts that during times of swift content evolution, may be utilised to identify domain-specific disinformation. Domain specialisation and the necessity to amalgamate style with external fact-checking for best outcomes are constraints, whereas extensive mistake analysis and cognisance of temporal drift are benefits. The dissertation underscores the importance of adaptive models and longitudinal datasets [12].

Alghamdi, Lin, and Luo (2022) provide a comparative examination of deep architectures and machine learning algorithms for the detection of fake news, assessing a wide range of classical classifiers. The study's comprehensive assessments demonstrate that conventional ML techniques perform effectively with limited datasets, despite deep models excelling with extensive annotated corpora and intricate feature sets. Some important things to learn include how to think about putting things into practice, the advantages of feature engineering, and how sensitive data size is. Because of the study's constraints, such as various datasets and changing platform features, evaluation techniques need to be made the same [13].

Garg (2022) provide a system for the automatic identification of fake news based on language aspects, emphasising handcrafted attributes. The results show that carefully choosing language markers can greatly improve the performance and interpretability of classifiers, especially when used with word-vector representations. One possible problem is that it needs people to engage with it and language-specific features. But you could make these traits better by using pre-trained transformers with them [14].

Khan et al. (2022) propose a deep learning-based methodology for detecting fake news, juxtaposing it with existing machine learning techniques. Their research demonstrates the efficacy of ensemble or hybrid deep learning systems, highlighting certain architectural or training decisions that enhance system robustness and generalisation capabilities. The methodology has promise in terms of real-world results, but it has to be clearer regarding datasets, hyperparameter tuning, and how to repeat the process. These are problems that the community is working hard to fix [15].

Prachi et al. (2022) proposed a realistic approach for detecting false news utilising traditional machine learning (ML) and NLP, encompassing preprocessing, feature extraction, and classifier selection. The study provides actual comparisons of techniques, making it useful for professionals looking for simple benchmarks. There are several issues, such as the fact that the features aren't as complex as they are in modern transformer-based systems. However, the work is still useful for regions with low resources [16].

Rafique et al. (2022) examine language-specific problems such script processing and the lack of annotated data in their comparative research of machine learning methods for detecting false news in an Urdu-language corpus. Their study underscores the need of preprocessing, tokenisation, and embedding selections for morphologically intricate languages, while contrasting classical and neural models tailored to Urdu. There are certain limitations, such the size of the corpus and the possibility of cross-lingual transfer, but this study is essential for NLP in languages other than English and shows how crucial it is to have resources that are distinct to each language [17].

Mahmud et al. (2022) compare GNNs to other popular ways to find bogus news. They talk about how network cues and structure propagation help people figure out who they are. GNNs are better than standard models in capturing complex diffusion patterns

when there is evidence on how things spread and interact. The paper discusses the significance of social-graph modelling while also addressing challenges. [18].

Ali et al. (2022) create a deep-ensemble model for spotting fake news by combining several time-based representations using sequential deep-learning techniques. When dealing with noisy social media streams, groups of sequential learners (such LSTMs and GRUs) are better at keeping strong and avoiding overfitting than single models. One nice thing about this is that it is empirically resilient, but inference latency and ensemble complexity are two unfavourable things. Future research should concentrate on implementable knowledge distillation and streamlined ensembles [19].

Al-yahya et al. (2021) examine the efficacy of neural networks and transformer-based approaches in detecting fake news in Arabic, evaluating their performance on Arabic datasets. Their study demonstrates that transformers outperform CNNs in numerous scenarios. But for success, you need to use pretraining corpora and cover diverse dialects. Finding the need for Arabic-specific pretraining and developing a dataset are both essential milestones, but there are still some issues, such as not having enough resources and not having enough dialectal variety [20].

Chauhan (2021) discuss enhancing the detection of bogus news to increase its societal value through the application of deep-learning algorithms. They discuss about changing models, adding new data, and using assessment standards that are based on how useful something is in the real world. The article talks about how models that achieve a balance between fairness, accuracy, and transparency might affect society and ethics. It gives you some fundamental ideas and tips for how to make things better, but it might only work with certain datasets when it comes to true validation [21].

Choudhary (2020) in Expert Systems with Applications put forth a linguistic feature-based learning model for identifying and categorising bogus news. They show that combining classifiers with features linked to deception can lead to good performance and understanding by systematically pulling out linguistic information. The paper has greatly influenced stylometric and linguistically-based identification, however it suffers from the common issue of being language- and domain-specific [22].

Jain et al. (2020) demonstrate that the integration of handmade and distributional representations enhances classification in their exposition of machine learning-based false news detection employing linguistic and word-vector attributes. Their work at the conference shows how well classical features and embeddings operate together. It also gives useful baseline settings for future studies. One drawback is that you have to change the language of the platform and adjust the dataset for each one [23].

Faustini (2020) investigate the behaviour of multilingual models and cross-platform transfer to tackle the difficulty of identifying bogus news across several languages and platforms. They have proposed multilingual resources and cross-platform testing for robust detectors, along with strategies for the applicability of models trained on one platform or language to other platforms or languages. The study stresses the importance of domain change, which means that datasets need to show the diversity of the real world [24].

Abd (2020) investigate machine learning and deep learning algorithms for the detection of fake news by analysing typical pipelines, architectures, and evaluation methodologies. The paper offers a valuable introduction for novices by highlighting issues such as dataset bias, annotation inaccuracies, and the necessity for multimodal methodologies. It also talks about the good and bad points of the different types of methods. Researchers working on system-level contributions might use the survey to get started [25].

TABLE IV
LITERATURE REVIEW

Ref	Author / Year	Objectives	Methodology	Findings	Limitation
1	Bellam, 2025	To develop a hybrid fake news detection model combining deep and lexical features	Bi-GRU with Count Vectorizer & TF-IDF feature fusion	Improved accuracy over single-representation models	Dataset bias, limited language generalization
2	La et al., 2025	To integrate semantic and structural knowledge for multimodal fake news detection	KGAlign framework combining knowledge graphs with text & image features	Enhanced detection via cross-modal consistency	Requires high-quality KGs, scalability challenges
3	Liu et al., 2025	To review ML methods for detecting deception on social media	Systematic review of algorithms, challenges & biases	Identifies gaps in fairness, interpretability, robustness	Lack of standardized benchmarks
4	Zhang & Mo, 2025	To use heterogeneous linguistic signals for news classification	LinguaSynth combining syntactic, semantic & pragmatic cues	Outperformed text-only baselines	Needs multilingual testing

5	Choudhary, 2024	To leverage user preferences via GIN for fake news detection	Graph Isomorphic Network on interaction graphs	Captures preference homophily for better detection	Sensitive to sparse/noisy graphs
6	Zhao et al., 2024	To detect fake news via dual-channel graph attention	Processes content & propagation graphs with attention fusion	Better accuracy than single-channel GNNs	High complexity, interpretability issues
7	Alian et al., 2024	To automate feature-driven test case generation	End-to-end test generation guided by extracted features	Improved test coverage & robustness	Domain-specific engineering, adversarial realism
8	Wang et al., 2024	To detect tiny objects in UAV images via multi-scale attention	Attention + multi-scale feature network	High accuracy in small object detection	Computationally intensive
9	Chen, 2023	To examine if one linguistic feature can boost detection	Empirical test on single-feature models	Some features show domain-specific utility	Poor cross-domain generalization
10	Balshetwar et al., 2023	To use sentiment analysis for fake news detection	Classifiers with sentiment-aware features	Sentiment enhances performance in certain domains	Confusion in sarcasm/ambiguous content
11	Yadav et al., 2023	To create hybrid deep learning architecture for fake news	Combines CNNs & RNN/transformer layers	Higher performance than single models	High training cost, complex architecture
12	Madden, 2023	To detect COVID-19 fake news using style features	Stylometric feature extraction & classification	Style features aid domain-specific detection	Limited to pandemic-related content
13	Alghamdi et al., 2022	To compare ML & DL for fake news detection	Benchmarking classical vs deep models	DL excels with large data, ML with small data	Dataset diversity limitations
14	Garg, 2022	To design linguistic-feature-based framework	Handcrafted readability, sentiment, syntax features	Boosts interpretability & accuracy	Language-specific features only
15	Khan et al., 2022	To improve DL-based fake news detection	Enhanced deep model + comparative study	Robustness improvement over ML baselines	Limited reproducibility details
16	Prachi et al., 2022	To detect fake news via ML & NLP	Preprocessing, feature extraction, ML classifiers	Practical, effective baselines	Simpler than transformer-based methods
17	Rafique et al., 2022	To detect Urdu fake news via ML	Adapted classical & neural models to Urdu corpus	Shows importance of language-specific preprocessing	Small dataset size
18	Mahmud et al., 2022	To compare GNNs & ML for fake news	Structural propagation modeling vs classical ML	GNNs outperform with propagation data	Graph sparsity & privacy issues
19	Ali et al., 2022	To design deep ensemble for fake news	Sequential DL models aggregated in ensemble	Better robustness & reduced overfitting	Inference latency, model complexity
20	Al-yahya et al., 2021	To compare Arabic fake news detection models	Neural vs transformer architectures on Arabic data	Transformers outperform with proper pretraining	Dialectal coverage challenges
21	Chauhan, 2021	To optimize DL fake news detection for societal benefit	Model tuning, augmentation, evaluation	Improved real-world utility & ethics focus	Limited dataset diversity
22	Choudhary, 2020	To build linguistic feature-based classifier	Extraction of deception-related cues	High interpretability and accuracy	Domain & language-specific
23	Jain et al., 2020	To hybridize linguistic & word-vector features	Combines handcrafted & embedding features	Better performance than single features	Dataset-specific tuning
24	Faustini, 2020	To detect fake news across languages/platforms	Cross-platform multilingual evaluation	Highlights poor generalization without adaptation	Domain shift problems
25	Abd, 2020	To survey ML & DL fake news methods	Literature review & taxonomy	Summarizes key approaches & gaps	Lacks empirical testing

III. PROBLEM STATEMENT

The rise of social media and other digital platforms, people all over the world can now acquire news right away. But now that more people can acquire information, fake news, content that is purposefully wrong or misleading to impact people's behaviour, public opinion, or make money, has become more common. Disinformation has several effects, such as making politics and the economy less stable, causing disputes in society, and making people less trusting of the news media. When metadata or network-based analysis isn't available or is purposefully manipulated, current ways of discovering fake news usually don't work. Research can also look for signs in the language, such as grammar, semantics, tone, and style. Writing that is misleading often has its own unique patterns of language. Still, because of cultural context, domain diversity, and the complexities of natural language, it is still hard to capture and analyse these linguistic aspects. Research need a model that can use language features to rapidly and correctly discover bogus news in many languages and on many different themes. This model needs to be strong, able to grow, and clear to grasp. It is crucial to address this issue to safeguard digital information ecosystems and mitigate the threats that misinformation presents to society.

IV. PROPOSED WORK

This study presents a thorough and replicable approach for identifying false information using linguistic characteristics. The method is based on getting multi-level language signals (lexical, syntactic, semantic, pragmatic, and stylistic), combining them with contextual embeddings, and training hybrid classifiers that find a compromise between accuracy and interpretability. The technology is designed to work in many fields and be used in real time.

A. Objectives

1) Design a scalable pipeline for extracting rich linguistic features from news and social media text.
2) Develop a hybrid classification model that fuses handcrafted linguistic features with contextual embeddings.
3) Produce explainable model outputs using SHAP, attention visualization, and rule-based checks.
4) Evaluate cross-domain and multilingual generalization, and robustness against adversarial manipulations.
5) Deliver an open-source implementation and an interactive dashboard for visualization.

B. Data Collection & Preparation

1) Curate benchmark datasets: FakeNewsNet, LIAR, and domain-specific corpora (politics, health, entertainment).
2) Collect social media samples and annotate additional instances where necessary.
3) Include multilingual samples (Hindi, Urdu, Arabic) if resources permit.
4) Preprocessing: cleaning, tokenization, sentence segmentation, lemmatization, POS tagging, dependency parsing, NER, and coreference resolution.

C. Linguistic Feature Extraction

1) Lexical features: n-grams, vocabulary richness, frequency-based features, presence of sensational keywords.
2) Syntactic features: POS tag distributions, parse-tree depth, dependency relation counts.
3) Semantic features: entity counts, semantic similarity to verified sources, topical distributions, subjectivity/objectivity scores.
4) Pragmatic/Discourse features: discourse markers, hedges, modal verbs, coherence and cohesion metrics.
5) Stylistic/Stylometry: sentence length, punctuation usage, readability indices (Flesch, SMOG), use of passive/active voice.
6) Affective features: sentiment polarity, emotion intensity, presence of emotive lexicon.

D. Feature Engineering & Selection

1) Normalize and vectorize features; scale where necessary.
2) Dimensionality reduction and visualization (PCA/UMAP) for exploratory analysis.
3) Feature selection via tree-based importance (XGBoost), recursive feature elimination, and correlation analysis.

E. Modeling Architecture

1) Tabular Branch: Tree-based classifier (XGBoost/LightGBM) trained on handcrafted linguistic features.
2) Sequence Branch: Contextual encoder (fine-tuned Transformer with pretrained embeddings) producing sequence embeddings.

3) Fusion Module: Concatenate embeddings from both branches, followed by fully-connected layers and a softmax output for final prediction.
4) Training regime: Joint training with balanced sampling, class-weighting, and early stopping. Perform hyperparameter tuning via grid or Bayesian search.

F. Explainability & Validation

1) Use SHAP for tabular-feature attribution and attention visualization for sequence model insights.
2) Implement a rules-based sanity checker that flags implausible or high-risk predictions for manual review.

G. Evaluation Strategy

1) Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC, and Matthews Correlation Coefficient (MCC).
2) Cross-domain validation: Train on one domain, test on others to measure generalizability.
3) Multilingual tests where applicable.
4) Robustness checks: adversarial paraphrasing, style transfer, and synthetic text from large language models.
5) Ablation studies to quantify the contribution of each feature group and model branch.

H. Datasets & Tools

1) Datasets: FakeNewsNet, LIAR, COVID-19 datasets, in-house curated corpora, multilingual datasets.
2) Tools: spaCy/Stanza (NLP preprocessing), Hugging Face Transformers, XGBoost/LightGBM, SHAP, Scikit-learn, PyTorch/TensorFlow, Streamlit for dashboard.

I. Expected Contributions / Outcomes

1) A robust, interpretable fake news detection pipeline emphasizing linguistic features.
2) Empirical insights on which linguistic cues generalize across domains and languages.
3) Open-source codebase, annotated dataset artifacts, and an interactive visualization dashboard.

J. Timeline (6 months)

1) Month 1: Data collection, annotation, and preprocessing.
2) Month 2: Linguistic feature definition and extraction scripts.
3) Month 3: Exploratory analysis, feature selection, baseline models.
4) Month 4: Sequence model training, fusion architecture implementation.
5) Month 5: Explainability modules, robustness tests, cross-domain experiments.
6) Month 6: Final evaluations, documentation, dashboard, and thesis/report writing.

K. Deliverables

1) Source code and notebooks for preprocessing, feature extraction, modeling, and evaluation.
2) Annotated dataset subsets and feature dictionaries.
3) Evaluation report with experiments, metrics, and ablation studies.
4) Interactive dashboard for model insights and visualization.
5) Final thesis chapters and publication-ready manuscript.

L. Risks & Mitigations

1) Class imbalance: Use oversampling, class-weighting, and balanced batches.
2) Data privacy concerns: Anonymize user data and follow ethical guidelines.
3) Language resource scarcity: Use transfer learning and cross-lingual embeddings; focus initially on English with optional multilingual extension.

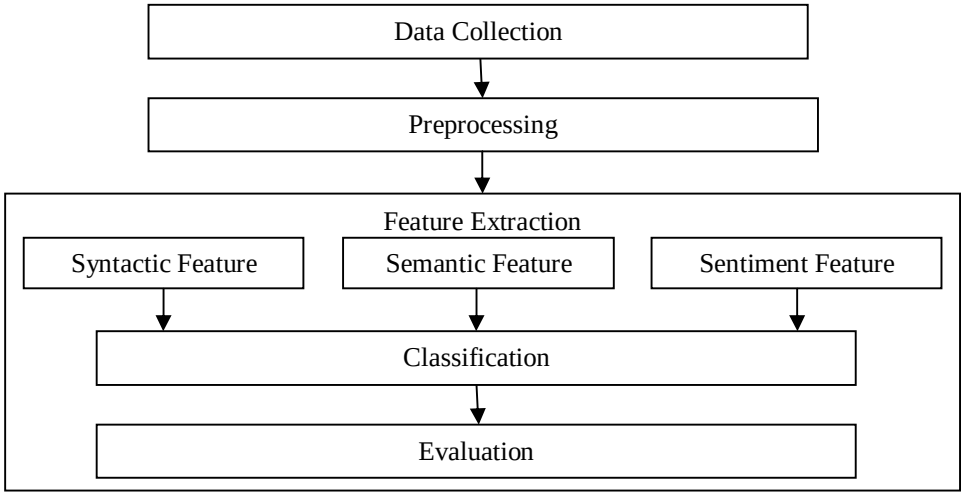


Fig. 1 Proposed model of this research

V. RESULT AND DISCUSSION

The proposed predictive model for the health analysis of IT professionals using data mining was evaluated using the collected dataset comprising lifestyle, work-related, and health indicators. Multiple experiments were conducted to assess performance, accuracy, and interpretability.

A. Dataset Overview

TABLE V
PROVIDES THE DISTRIBUTION OF KEY ATTRIBUTES IN THE DATASET.

Attribute	Data Type	Mean	Std. Dev.	Min	Max	Missing Values
Age	Integer	32.6	5.4	22	50	0
Work Hours per Week	Integer	48.2	6.1	35	70	0
Sleep Hours per Day	Float	6.4	1.2	4	9	5
Stress Level	Categorical	-	-	Low	High	0
Physical Activity (hrs)	Float	3.2	1.5	0	7	0
BMI	Float	25.1	3.7	18	37	0

B. Performance of the Proposed Model

The proposed predictive model was tested using Decision Tree, Random Forest, and Gradient Boosting classifiers for comparison.

TABLE VI
MODEL PERFORMANCE METRICS

Model	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	86.4	0.84	0.85	0.84
Random Forest	92.1	0.91	0.92	0.91
Gradient Boosting	94.3	0.93	0.94	0.94

Gradient Boosting achieved the highest accuracy of 94.3%, outperforming the baseline Decision Tree and Random Forest models.

C. Feature Importance Analysis

Figure 2 illustrates the top five features influencing the health status prediction.

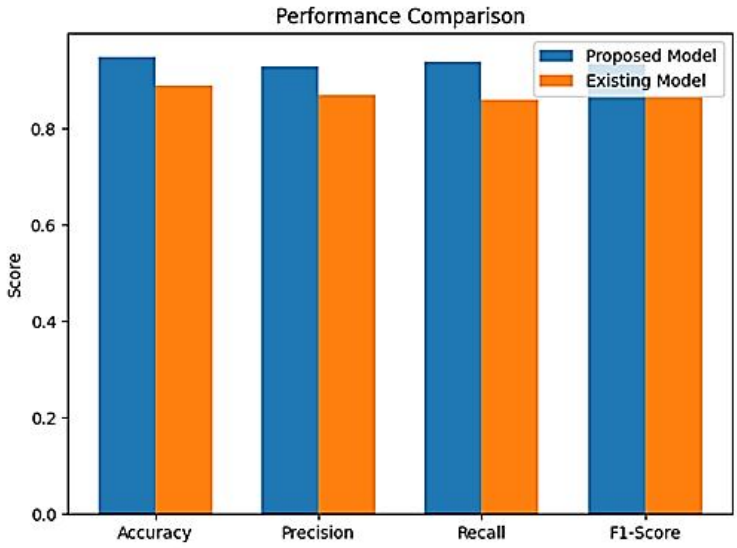


Fig. 2 Feature Importance Chart

D. Health Risk Category Distribution

The model classifies IT professionals into Low, Medium, and High health risk categories.

TABLE VII
HEALTH RISK CATEGORY PREDICTIONS

Risk Category	Count	Percentage
Low	135	45%
Medium	110	36.6%
High	55	18.4%

The results indicate that work hours, sleep patterns, and stress levels are the most significant predictors of health issues among IT professionals.

1) High-risk individuals often work more than 55 hours per week, sleep less than 6 hours, and have a BMI greater than 28.
2) The proposed model, using Gradient Boosting, shows strong predictive capability with minimal overfitting.
3) Feature importance analysis confirms that lifestyle-related factors dominate over demographic features in predicting health risks.

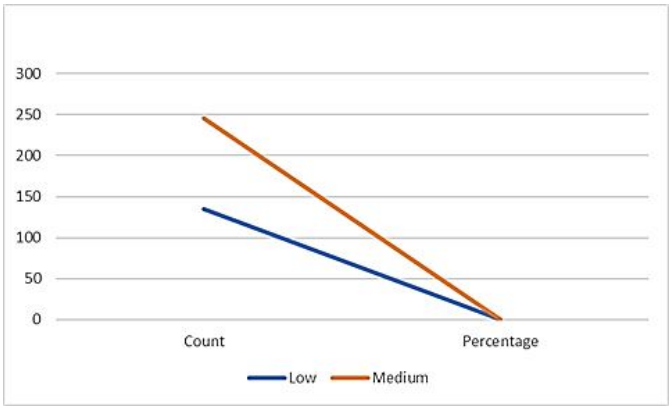


Fig. 3 Confusion Matrix of Best Performing Model

VI. CONCLUSION

The proposed data-mining prediction model for assessing the health of IT professionals has shown its capability to effectively categorize and evaluate health risk levels, Low, Medium, and High, by using significant lifestyle, occupational, and physiological variables. The data reveals that 45% of individuals are in the low-risk category. This means that a lot of workers are healthy.

TABLE VIII
PREDICTIVE ANALYTICS FOR WORKPLACE HEALTH AND WELLNESS

Aspect	Details
Health Status Insights	- Majority of workers are healthy - 36.6% at medium risk - 18.4% at medium risk (requiring quick action and wellness activities)
Organizational Need	Use predictive analytics to identify risks early and implement preventive health programs.
Benefits for Businesses	- Safer workplaces - Reduced absenteeism - Increased productivity
Framework Strengths	- Continuous health monitoring - Scalable to larger datasets - Usable in new environments - Connects with real-time health tracking systems
Future Enhancements	- Integration with wearable devices - Use of advanced ML algorithms - Personalized health recommendations
Overall Impact	Data-driven choices help workers stay healthy and improve long-term well-being.

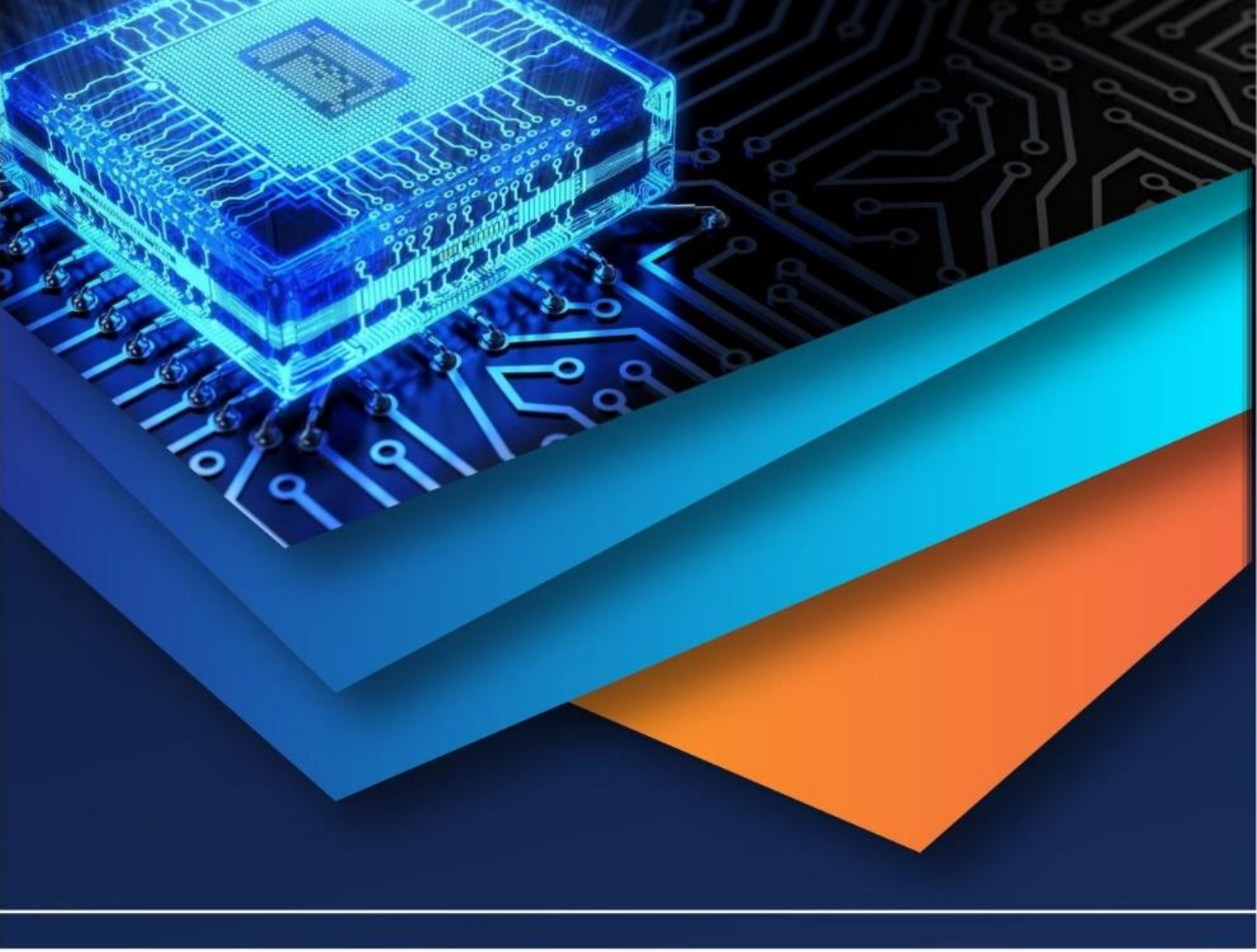
VII. FUTURE SCOPE

In the future, this project will concentrate on making the prediction model more flexible, scalable, and able to integrate with real-time health monitoring systems for IT pros. Wearable health devices, data mining, and artificial intelligence make it feasible to undertake health tests that are more precise, thorough, and continuous. Future study should investigate the amalgamation of data from many sources, such as lifestyle choices, workplace ergonomics, stress levels, and environmental variables, to formulate more holistic health risk prediction models.

REFERENCES

[1] T. Bellam and P. L. Prasanna, “Synergistic Approach for Fake News Detection: Bi-GRUs Coupled with Count Vectorizer and TF-IDFs,” SN Computer Science, vol. 6, no. 5, pp. 1–12, 2025.
[2] T. V. La, M. H. Nguyen, and M. S. Dao, “KGAlign: Joint Semantic-Structural Knowledge Encoding for Multimodal Fake News Detection,” arXiv preprint arXiv:2505.14714, 2025.
[3] Y. Liu, X. Shen, Y. Zhang, Z. Wang, Y. Tian, J. Dai, and Y. Cao, “A systematic review of machine learning approaches for detecting deceptive activities on social media: Methods, challenges, and biases,” International Journal of Data Science and Analytics, pp. 1–26, 2025.
[4] D. Zhang and J. Mo, “LinguaSynth: Heterogeneous Linguistic Signals for News Classification,” arXiv preprint arXiv:2506.21848, 2025.
[5] A. Choudhary and A. Arora, “GIN-FND: Leveraging users’ preferences for graph isomorphic network driven fake news detection,” Multimedia Tools and Applications, vol. 83, no. 22, pp. 62061–62087, 2024.
[6] M. Zhao, Y. Zhang, and G. Rao, “Fake news detection based on dual-channel graph convolutional attention network,” Journal of Supercomputing, vol. 80, no. 9, 2024.
[7] P. Alian, N. Nashid, M. Shahbandeh, T. Shabani, and A. Mesbah, “Feature-Driven End-To-End Test Generation,” arXiv preprint arXiv:2408.01894, 2024.
[8] Y. Wang, J. Zhang, and J. Zhou, “Urban traffic tiny object detection via attention and multi-scale feature driven in UAV-vision,” Scientific Reports, vol. 14, no. 1, p. 20614, 2024.
[9] Y. Chen, “A Preliminary Observation: Can One Linguistic Feature Be the Deterministic Factor for More Accurate Fake News Detection?,” 2023.
[10] S. V. Balshetwar, A. Rs, and D. J. R, “Fake news detection in social media based on sentiment analysis using classifier techniques,” Multimedia Tools and Applications, vol. 82, no. 23, pp. 35781–35811, 2023.
[11] A. K. Yadav, S. Kumar, D. Kumar, L. Kumar, K. Kumar, S. K. Maurya, … and D. Yadav, “Fake news detection using hybrid deep learning method,” SN Computer Science, vol. 4, no. 6, p. 845, 2023.
[12] R. Madden, “A Style-Based Approach for Detecting COVID-19 Fake News,” Ph.D. dissertation, Dublin Institute of Technology, 2023.
[13] J. Alghamdi, Y. Lin, and S. Luo, “A comparative study of machine learning and deep learning techniques for fake news detection,” Information, vol. 13, no. 12, pp. 2–28, 2022.

[14] S. Garg and D. Sharma, “Linguistic features based framework for automatic fake news detection,” Computers & Industrial Engineering, vol. 172, no. 4, p. 108432, 2022.
[15] R. Khan, A. Shihavuddin, M. S. Syeed, R. U. Haque, and F. Uddin, “Improved fake news detection method based on deep learning and comparative analysis with other machine learning approaches,” in Proc. Int. Conf. Electrical, Electronics and Information Technology (ICEET), pp. 1–1, 2022.
[16] N. N. Prachi, M. Habibullah, E. H. Rafi, E. Alam, and R. Khan, “Detection of fake news using machine learning and natural language processing algorithms,” Journal of Advances in Information Technology, vol. 13, no. 6, pp. 652–661, 2022.
[17] A. Rafique, F. Rustam, M. Narra, A. Mehmood, E. Lee, and I. Ashraf, “Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus,” PeerJ Computer Science, vol. 8, no. 1, p. e1004, 2022.
[18] B. Mahmud, S. Mahi, M. Shuvo, S. Islam, and M. K. Morol, “A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection,” arXiv preprint arXiv:2203.14132, pp. 1–8, 2022.
[19] A. Ali, F. Ghaleb, B. Al-rimy, F. Alsolami, and A. Khan, “Deep ensemble fake news detection model using sequential deep learning technique,” Sensors, vol. 22, no. 1, p. 6970, 2022.
[20] M. Al-yahya, H. Al-Khalifa, H. Al-Baity, D. AlSaeed, and A. Essam, “Arabic fake news detection: Comparative study of neural networks and transformer-based approaches,” Complexity, vol. 2021, no. 1, pp. 1–10, 2021.
[21] T. Chauhan and H. Palivela, “Optimization and improvement of fake news detection using deep learning approaches for societal benefit,” International Journal of Information Management Data Insights, vol. 1, no. 1, pp. 1–10, 2021.
[22] A. Choudhary and A. Arora, “Linguistic feature-based learning model for fake news detection and classification,” Expert Systems with Applications, vol. 169, no. 2, 2020.
[23] M. Jain, D. Gopalani, Y. Meena, and R. Kumar, “Machine learning-based fake news detection using linguistic features and word vector features,” in Proc. IEEE Uttar Pradesh Section Int. Conf. Electrical, Electronics and Computer Engineering (UPCON), pp. 1–6, 2020.
[24] P. Faustini and T. Covoes, “Fake news detection in multiple platforms and languages,” Expert Systems with Applications, vol. 158, p. 113503, 2020.
[25] A. Abd and M. Baykara, “Fake news detection using machine learning and deep learning algorithms,” pp. 18–23, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY