IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

AI-Powered Prediction System Using NLP and Deep Learning

Dr. Pankaj Agarkar[1], Tushar Zalte[2] , Farhan Shikalgar[3], Umesh Kardile[4]
Department of Computer Engineering
Ajeenkya DY Patil School of Engineering Lohegaon, Pune, India.

Abstract: *Prediction systems grounded in textual data have become indispensable across high-stakes domains including clinical decision support, financial signal detection, and digital misinformation analysis. Classical statistical approaches and shallow machine learning methods have demonstrated satisfactory performance on narrow, well-curated datasets, but they struggle to generalise once input distributions shift or domain vocabulary diverges from training corpora. Deep learning, and more specifically the pre-trained transformer paradigm, has substantially narrowed this gap; nevertheless, single-architecture solutions routinely leave accuracy on the table when applied to tasks that demand both rich contextual encoding and explicit sequential reasoning.*

This paper presents a cohesive, end-to-end AI- powered prediction framework that fuses BERT- derived contextual representations with a two-layer bidirectional LSTM (BiLSTM) classification head augmented by an additive attention mechanism. The system is designed as a modular pipeline: text acquisition and normalisation, augmentation-based imbalance handling, deep encoding, sequential modelling, and post-hoc probability calibration are treated as independent, replaceable stages. Experimental evaluation across three publicly available benchmark datasets — the LIAR fake news corpus, Stanford Sentiment Treebank v2, and a health-claim verification collection — confirms that the hybrid BERT-BiLSTM-Attention architecture outperforms five competitive baselines on macro-averaged F1 and area under the ROC curve. Ablation experiments quantify the individual contributions of the attention layer, recurrent head, augmentation strategy, and temperature scaling. A discussion of deployment trade-offs addresses inference latency, continual adaptation, and algorithmic fairness..

Keywords: Natural Language Processing, Deep Learning, BERT, Bidirectional LSTM, Attention Mechanism, Transfer Learning, Intelligent System Predictive Modelling, Text Classification, Calibration

I. INTRODUCTION

Language has always been humanity's primary vehicle for recording observation, intention, and judgment. As organisations accumulate text at unprecedented scale — spanning electronic health records, regulatory filings, social media streams, and internal communications — the prospect of deriving reliable predictive signals from this content has grown from a theoretical curiosity into an operational priority. Prediction in this context does not merely mean retrieval or summarisation; it means inferring the future value of a variable, or the truth status of a claim, from textual evidence that was not explicitly curated for that purpose. Meeting this challenge requires methods that are sensitive to syntax, semantics, pragmatics, and the statistical regularities that emerge across large corpora[1].

The journey toward language-aware prediction has unfolded in distinct waves. The first wave relied on manually engineered features: keyword counts, part-of-speech ratios, readability scores, and domain-specific lexicons. While these features embodied hard-won linguistic intuition, they proved brittle against vocabulary drift and transferred poorly across domains. The second wave introduced data-driven representations through neural word embeddings, which compressed vocabulary into dense vectors whose geometric relationships encoded semantic kinship. Feedforward and recurrent classifiers operating on these embeddings achieved new benchmarks on sentiment analysis, question classification, and intent detection.[2]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

A third wave, still unfolding, is defined by pre- trained transformer language models. Rather than embedding individual words, transformers produce representations conditioned on entire surrounding contexts, resolving the polysemy that crippled context-free embeddings. Pre-training on web-scale corpora endows these models with broad linguistic knowledge that can be transferred to downstream tasks through fine-tuning on comparatively modest labelled datasets. BERT and its successors demonstrated that this transfer learning recipe generalises across an extraordinary range of NLP benchmarks[3].

Linguistic understanding can be effectively transferred to downstream applications by fine-tuning models on relatively small labeled datasets. BERT and the models that followed showed that this transfer learning approach works successfully across a wide variety of NLP tasks and benchmarks[4].

Yet even within the transformer paradigm, design choices matter. A BERT encoder fine-tuned with a simple linear head is computationally efficient and often highly competitive, but it encodes sequential ordering implicitly through fixed positional embeddings rather than through an architectural commitment to order. For tasks where the linear arrangement of propositions, hedges, and logical connectives carries predictive weight — fake news classification, clinical outcome prediction, legal document analysis — there is reason to supplement the transformer encoder with a recurrent layer that treats order as a first-class modelling concern[5].

This paper investigates the hypothesis that routing BERT's contextual representations through a bidirectional LSTM, further equipped with an additive attention mechanism, yields consistent accuracy improvements over standalone architectures, particularly on tasks characterised by low resource availability, class imbalance, or complex argumentative structure. We situate this investigation within a production-oriented system design that addresses not only model accuracy but also calibration quality, inference cost, fairness, and interpretability[6].

1.1 Motivation and Problem Framing

Three application scenarios inform the design of this system. In the domain of news veracity assessment, a prediction system must determine whether a textual claim is supported by established facts, a task that requires sensitivity to linguistic markers of certainty, source attribution, and logical consistency across sentences. In clinical risk stratification, free-text nursing notes and physician assessments contain nuanced hedges and qualifiers whose sequential arrangement influences outcome likelihood. In customer experience management, contact centre transcripts and survey responses must be mapped to actionable outcome categories to support operational decisions[7].

What unites these scenarios is their shared reliance on text as the primary data modality, combined with the requirement for well-calibrated probability estimates rather than simple hard classifications. A model that correctly ranks a positive example above a negative one but assigns a probability of 0.98 to every positive case is poorly calibrated and unsuitable for downstream decision-making frameworks that threshold on predicted risk. System design must therefore address calibration explicitly rather than treating it as an afterthought.

1.2 Research Contributions

The specific contributions of this paper are as follows. First, we describe a complete, modular prediction pipeline that progresses from raw text ingestion through preprocessing, augmentation, deep encoding, sequential classification, and calibration, with each stage designed for independent replacement. Second, we introduce and evaluate a hybrid BERT-BiLSTM-Attention model as the central inference component and compare it against five baselines under identical experimental conditions across three benchmark datasets. Third, we present an ablation study that isolates the contribution of each pipeline component, providing evidence-based guidance for practitioners who must balance accuracy against computational cost. Fourth, we offer an extended discussion of deployment considerations including latency management, domain adaptation strategies, fairness measurement, and interpretability techniques appropriate for production environments[8].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

1.3 Paper Organisation

The remainder of the paper is structured as follows. Section 2 surveys relevant prior work on text representations, transformer architectures, recurrent and hybrid models, and existing prediction systems. Section 3 details the proposed system architecture and each pipeline stage. Section 4 describes implementation specifics, training procedures, and evaluation protocols. Section 5 presents experimental results, comparative analysis, and ablation findings. Section 6 discusses practical deployment considerations. Section 7 concludes the paper and identifies directions for future research[9].

II. BACKGROUND AND RELATED WORK

2.1 From Bag-of-Words to Dense Representations

The bag-of-words model treats a document as an unordered multiset of tokens, discarding sequence structure entirely. Its appeal lies in computational simplicity: documents become sparse vectors in a vocabulary- dimensional space, and linear classifiers trained on these vectors achieve competitive performance on tasks where word co-occurrence patterns carry most of the predictive signal. Term frequency-inverse document frequency weighting dampens the influence of terms that appear frequently across all documents, improving precision but not addressing the fundamental ordering problem[10].

Latent semantic analysis introduced the idea of recovering latent semantic dimensions from a term-document co-occurrence matrix through truncated singular value decomposition, producing lower-dimensional document representations that grouped semantically related content[11].

Probabilistic latent semantic analysis and Latent Dirichlet Allocation subsequently cast document modelling as a generative process involving mixtures of topics, yielding interpretable topic distributions that could serve as classification features. Both families remain useful in low- resource settings but lack the expressiveness of neural representations[12].

Word2Vec introduced the continuous bag-of-words and skip-gram architectures, training shallow neural networks to predict words from their context or contexts from a word, respectively[13]. The resulting dense word vectors exhibited remarkable algebraic structure: vector arithmetic could recover analogical relationships. GloVe augmented the skip-gram objective with a global co-occurrence constraint, and FastText extended both approaches to character n-gram subword representations, enabling reasonable vectors for rare and morphologically rich words[14].

2.2 Sequential Models for Text

Recurrent neural networks process sequences token by token, maintaining a hidden state that theoretically encodes all prior context. In practice, standard recurrent networks suffer from vanishing gradients that prevent effective learning of dependencies spanning more than a handful of positions[15]. Long Short-Term Memory networks addressed this limitation through three multiplicative gates — input, forget, and output — that control the flow of information into, through, and out of a memory cell. The forget gate in particular allows the network to selectively discard irrelevant context, effectively maintaining information over arbitrarily long spans when the task demands it[16].

Gated Recurrent Units offer a simplified gating mechanism that merges the input and forget gates into a single update gate and eliminates the separate cell state, yielding a computationally lighter alternative with broadly comparable performance. Bidirectional variants of both LSTM and GRU process sequences in both the forward and backward directions, allowing each position's representation to incorporate global context from the full sequence, which is particularly beneficial for classification tasks where evidence may appear anywhere in the input[17].

Convolutional neural networks adapted for text apply sets of one-dimensional filters of varying widths to the token sequence, detecting local n-gram patterns with shift- invariance. While lacking the long-range modelling capacity of recurrent networks, convolutional architectures are highly parallelisable and empirically competitive on tasks dominated by local lexical and syntactic patterns. Architectures combining both convolutional and recurrent layers have been explored to combine the efficiency of local feature detection with the sequential coherence of recurrence[18].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

2.3 Transformer Architecture and Pre-Training

The transformer architecture replaced recurrence entirely with multi-head self-attention, allowing every position in a sequence to directly attend to every other position in a single layer. Scaled dot-product attention computes compatibility scores between query and key representations, normalises them via softmax, and uses the resulting weights to form a value-weighted output. Multi- head attention applies this mechanism independently in multiple representation subspaces and concatenates the outputs, enabling the model to jointly attend to information from different positions and representation types[19].

BERT adapted the encoder stack of the transformer for self-supervised pre-training on two objectives: masked language modelling, where 15% of tokens are replaced with a mask token and the model must predict them from bidirectional context, and next-sentence prediction, where the model classifies whether two sentences are consecutive in the original document. Pre-training on the BooksCorpus and English Wikipedia yielded a 12-layer base model and a 24- layer large model whose representations transferred effectively to a wide array of downstream tasks through lightweight fine-tuning[20].

Subsequent BERT variants refined the pre-training recipe in various ways. RoBERTa trained for longer on more data without next-sentence prediction, demonstrating that training duration and data scale were underexploited in the original BERT release. ALBERT introduced factored embedding parameterisation and cross-layer parameter sharing to reduce model size without proportional performance loss. DistilBERT applied knowledge distillation to compress a 12-layer BERT into a 6-layer student that retained approximately 97% of performance at 60% of the parameter count, a trade-off attractive for deployment- constrained settings[21].

Domain-specific variants demonstrated the importance of pre-training corpus composition. BioBERT, pre-trained on PubMed abstracts and PubMed Central full- text articles, outperformed general BERT on biomedical named entity recognition and relation extraction. FinBERT, pre-trained on financial communications, improved sentiment analysis of earnings call transcripts. These findings suggest that continued pre-training on domain text, even for a relatively small number of steps, substantially narrows the distribution gap between general and target vocabularies[22].

2.4 Attention Mechanisms and Their Roles

The term attention covers a family of mechanisms that all serve the same purpose: computing a context- dependent weighted combination of a set of input vectors. The original Bahdanau additive attention was introduced for sequence-to-sequence models to allow the decoder to selectively focus on different encoder positions when generating each output token. In classification settings, a similar mechanism can aggregate a variable-length sequence of encoder outputs into a fixed-length vector by assigning learned alignment scores to each position[23].

The relationship between transformer self-attention and classification-specific attention is worth clarifying. Self-attention within the transformer encoder is a many-to-many operation that updates every token's representation by attending to all others; it operates within the encoding stage. Classification attention, as used in this work, is a many-to-one aggregation applied after the BiLSTM, compressing the sequence of position-wise hidden states into a single summary vector for the classification head. Both mechanisms are present in the proposed architecture and serve complementary roles[24].

Attention weights have been frequently cited as a source of model interpretability, and this claim deserves nuanced treatment. Attention weights reflect the model's internal routing of information, but they are not guaranteed to correspond to the human notion of importance. Gradient- based attribution methods, which measure how much a small perturbation of an input token changes the predicted output, provide a complementary and theoretically better grounded importance signal. Practical interpretability tooling ideally surfaces both perspectives[25].

2.5 Related Prediction Systems

The fake news detection literature provides a direct antecedent for the proposed system. Early approaches extracted stylistic signals — hedging language frequency, sentiment polarity, vocabulary richness, and readability indices — and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

trained linear classifiers on these hand-crafted features. Later work leveraged graph neural networks to model propagation patterns on social media platforms, treating the topology of retweet and reply networks as a supplementary evidence channel. Transformer-based approaches have achieved strong benchmark performance when focusing on claim-level linguistic signals, though combining linguistic and graph-based evidence remains an active research direction[26].

Financial NLP prediction systems face distinct challenges. Earnings call transcripts mix structured financial reporting with qualitative management commentary, and the predictive signal for stock movement resides disproportionately in cautious or optimistic linguistic framing rather than in the reported numbers themselves. Models trained on these transcripts must handle domain-specific jargon, forward-looking statements with regulatory caveats, and deliberate ambiguity introduced by management communication strategy[27].

Clinical NLP predictive systems have demonstrated measurable value in tasks such as hospital readmission prediction, in-hospital mortality risk stratification, and adverse drug event detection. A recurring finding in this literature is that free-text clinical notes contain prognostic information not captured by structured electronic health record variables, motivating the inclusion of NLP-derived features alongside laboratory values and diagnosis codes in multimodal prediction models. Privacy considerations and regulatory constraints introduce deployment challenges not faced in other domains, underscoring the importance of on- premises deployment options and rigorous de-identification pipelines[28].

III. PROPOSED SYSTEM ARCHITECTURE

3.1 Architectural Overview

The proposed system is organised as a five-stage pipeline: (1) data acquisition and ingestion, (2) preprocessing and augmentation, (3) contextual encoding via a pre-trained BERT model, (4) sequential classification via a BiLSTM with additive attention, and (5) output calibration and serving. Each stage communicates with adjacent stages through a well-defined internal schema, enabling components to be upgraded or replaced without global code changes. This separation reflects the practical reality that different deployment scenarios — high-throughput batch processing versus low-latency online inference — demand different trade-offs at specific stages rather than architectural overhauls[29].

The design philosophy prioritises modularity over monolithic efficiency. A unified codebase that tightly couples the encoder and classifier would be marginally faster in training due to reduced inter-component data transfer, but it would severely limit the ability to swap encoder backbones as better pre-trained models become available. Given the pace of progress in pre-trained language modelling, architectural flexibility is a more durable engineering investment than marginal training speed gains[30].

3.2 Data Acquisition and Ingestion

The ingestion layer accepts text data from three source types: relational databases accessed via parameterised SQL queries, file system directories containing documents in plaintext, HTML, PDF, or JSON format, and streaming APIs that deliver documents through webhook or long-polling protocols. A format-aware parser normalises each source into a canonical internal document object containing a unique identifier, a raw text field, an optional label, a timestamp, and a key-value metadata dictionary[31].

For streaming sources, a lightweight buffer with configurable flush interval accumulates documents into mini- batches before forwarding them to the preprocessing stage. Backpressure handling prevents memory overflow when downstream processing is temporarily slower than the ingestion rate. Duplicate detection via locality-sensitive hashing discards near-identical documents before they enter the preprocessing stage, preventing duplicate examples from inflating training set statistics[32].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

3.3 Preprocessing and Text Normalisation

Raw text ingested from heterogeneous sources exhibits considerable variation in encoding, formatting, and lexical noise. The preprocessing module applies a deterministic sequence of transformations designed to reduce spurious variation while preserving semantically relevant signal. Unicode normalisation under the NFKC form collapses visually equivalent characters to a canonical representation, preventing the same word from appearing as distinct types due to encoding accidents. HTML and XML markup is stripped using a tag-aware parser that preserves embedded text content; a naïve regular expression approach is avoided because it fails on malformed markup common in web-scraped corpora[33].

URLs and email addresses are replaced with category tokens — [URL] and [EMAIL] respectively — rather than removed entirely, preserving their positional and structural role in the text while eliminating vocabulary sparsity introduced by unique addresses. Numeric sequences longer than six digits, which typically encode identifiers rather than quantitative values, are similarly replaced with an [ID] token. Social media text receives additional normalisation: repeated characters in elongated spellings are collapsed to a maximum of two consecutive identical characters, and hashtag boundaries are segmented using a frequency-based word splitter[34].

Sentence segmentation applies a rule-augmented statistical boundary detector trained on newswire and web text, with additional heuristics for handling abbreviations, decimal numbers, and ellipses that commonly mislead simpler period-based splitters. For inputs longer than the BERT maximum sequence length, segmentation also enables the sliding-window strategy described in Section 3.4.

3.3.1 Data Augmentation for Class Imbalance

Class imbalance is endemic to real-world labelled text corpora: misinformation may represent a small fraction of all claims, rare adverse clinical outcomes may affect a minority of patients, and specific sentiment categories may be underrepresented in organic feedback data. Training directly on imbalanced distributions causes classifiers to optimise for majority-class accuracy at the expense of minority-class recall, producing systems that are statistically accurate but operationally unreliable for the cases where prediction matters most[35].

The system offers three augmentation strategies selectable per-dataset. Back-translation augmentation passes minority-class sentences through a neural machine translation model to an intermediate language — German or French, selected based on translation quality for the domain — and translates back to English. The round-trip introduces lexical and syntactic paraphrasing variation while preserving the core semantic content and label validity. Synonym substitution draws replacements from a distributional thesaurus, replacing randomly selected non-stopword tokens with contextually appropriate synonyms at a configurable substitution rate; overly aggressive substitution risks corrupting task-relevant signals and is therefore bounded at 20% of tokens per example. A weighted mini-batch sampler oversamples minority class indices during training without duplicating examples, ensuring that gradient updates reflect a more balanced label distribution[36].

3.4 Contextual Encoding

The encoding stage maps cleaned, normalised text into dense contextual representations using a pre-trained BERT model. WordPiece tokenisation splits each word into the longest subword units present in the pre-trained vocabulary, allowing rare and morphologically complex words to be represented through compositions of more frequent subword pieces. A special [CLS] token is prepended to each input and a [SEP] token appended; the [CLS] position's final hidden state serves as a pooled sequence representation under BERT's original formulation, though this paper supplements it with richer extraction strategies[37].

For inputs exceeding the maximum sequence length — 512 sub-word tokens for BERT-base — the sliding- window approach divides the input into overlapping windows of fixed length with a configurable stride. Each window is encoded independently, producing a set of token-level hidden state sequences. Representations for tokens that fall within the overlap region of multiple windows are averaged across those windows, preventing boundary artefacts that

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

would arise from hard truncation. The resulting full-sequence token representations are then passed to the BiLSTM stage[38].

Rather than extracting only the final transformer layer's hidden states, the system concatenates the hidden states from the last four transformer layers. This multi-layer extraction exploits the observation that different encoder layers encode different types of linguistic knowledge: lower layers capture morphological and syntactic properties while higher layers encode more abstract semantic content. For tasks that benefit from syntactic awareness — such as detecting logical connectives in argumentation — multi-layer concatenation consistently outperforms single-layer extraction[39].

3.5 BiLSTM Classification Head with Additive Attention

The sequence of per-token BERT representations is fed as input to a two-layer bidirectional LSTM. Each LSTM layer processes the sequence twice, once in the forward direction from the first token to the last, and once in the backward direction from the last token to the first. At each position, the forward and backward hidden states are concatenated to produce a joint representation that encodes both left and right context for that position explicitly. The two-layer stacking allows the second layer to operate on representations that already encode local context from the first layer, supporting the capture of longer-range dependencies[40].

After the second BiLSTM layer, an additive attention mechanism computes a scalar alignment score for each position by applying a learned linear transformation to the hidden state, projecting to a single dimension, and applying a hyperbolic tangent non-linearity followed by another linear projection. These raw scores are normalised across positions with a softmax function, yielding a probability distribution over the sequence. The context vector is computed as the weighted sum of the BiLSTM hidden states under this distribution, concentrating representational mass on the positions the model has learned to associate most strongly with the target label[41].

The context vector is regularised with dropout applied at a tunable retention probability, a standard stochastic regularisation technique that prevents co- adaptation of hidden units by randomly zeroing a subset of activations during training. The dropped representation is passed through a linear projection layer whose output dimensionality equals the number of target classes. For binary tasks a sigmoid activation converts the scalar output to a probability, and the binary cross-entropy loss is used during training. For multi-class tasks a softmax activation produces a valid probability distribution and the categorical cross- entropy loss is applied[42].

3.6 Post-Training Probability Calibration

Neural classifiers trained with cross-entropy loss on large parameter spaces are notorious for overconfidence: the predicted probability mass concentrates near zero and one even when the model's effective discrimination capability does not justify such certainty. This overconfidence is particularly pronounced after fine-tuning from a pre-trained initialisation, where the training procedure may converge to sharp minima in the output logit space[43].

Temperature scaling addresses this problem by introducing a single positive scalar parameter T, the temperature, and dividing the pre-softmax logits by T before the final activation. When T is greater than one, the logit distribution is flattened, reducing overconfidence; when T is less than one, predictions become sharper. The temperature is optimised on a held-out calibration partition by minimising the negative log-likelihood of the calibrated predictions; because T affects only the output probabilities and not the argmax, it does not alter class assignments. Temperature scaling is computationally negligible and does not require retraining the underlying model, making it a highly practical post-processing step[44].

Expected Calibration Error (ECE) measures calibration quality by partitioning predicted probabilities into bins, computing the gap between average predicted probability and average empirical accuracy within each bin, and taking a weighted average of these gaps. A perfectly calibrated model would produce an ECE of zero; typical uncalibrated deep models exhibit ECE values in the range of 0.05–0.15. After temperature scaling in the proposed system, ECE consistently falls below 0.03 across all three benchmark datasets[45].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

IV. IMPLEMENTATION DETAILS

4.1 Software Stack and Dependencies

The system is implemented entirely in Python 3.10. PyTorch 2.0 provides the automatic differentiation engine, GPU memory management, and the core neural network module abstractions. The Hugging Face Transformers library (version 4.36) supplies pre-trained model weights, tokenisers, and the model configuration schemas needed for reproducible checkpoint loading. Datasets from the Hugging Face Hub are loaded via the datasets library, which supports efficient memory-mapped storage for large corpora and streaming loading for datasets that exceed available RAM[46].

Custom PyTorch Dataset and DataLoader subclasses manage the tokenisation, padding, and label encoding steps at the sample and batch levels respectively. The DataLoader's collate function handles variable-length sequence batching through dynamic padding to the maximum length within each batch rather than the global maximum, reducing average computation by approximately 30% on corpora with high sequence length variance. The weighted sampler for class imbalance correction is implemented as a custom PyTorch Sampler that computes per-sample weights inversely proportional to class frequency[47].

Experiment configuration is managed through a YAML-based configuration system that records all hyperparameters, dataset paths, random seeds, and model checkpoint identifiers. Training runs are logged to a local SQLite database, with each run assigned a unique hash-based identifier derived from its configuration. This design supports exact reproducibility: given a configuration identifier, the system can reconstruct the exact training conditions of any prior run[48].

4.2 Model Configuration and Hyperparameters

Pre-trained weights are loaded from the bert-base- uncased checkpoint, which encodes 12 transformer layers, 768 hidden dimensions per layer, 12 attention heads, a feed- forward intermediate dimension of 3072, and approximately 110 million parameters in total. The base variant is preferred over bert-large for the experiments reported here because it delivers a favourable accuracy-to-latency trade-off; preliminary experiments confirmed that bert-large yielded less than 0.5 points of additional F1 at more than twice the inference cost.

The BiLSTM is configured with a hidden size of 256 per direction, yielding a 512-dimensional concatenated hidden state at each position. Two recurrent layers are stacked, with inter-layer dropout applied at a rate of 0.1 to provide mild regularisation without impairing gradient flow. The attention projection dimension is set to 128. The final linear classifier applies dropout at a rate of 0.3 before the projection, a value selected on the validation set via a coarse grid search over {0.1, 0.2, 0.3, 0.4}.

Fine-tuning uses the AdamW optimiser with a peak learning rate of 2e-5 for the BERT encoder parameters and 1e-3 for the freshly initialised BiLSTM and attention parameters, reflecting the difference in initialisation quality between the pre-trained and randomly initialised components. A linear learning rate schedule with 6% of total training steps dedicated to warm-up prevents destructive updates to the pre- trained representations during the early fine-tuning phase. The maximum gradient norm is clipped to 1.0 throughout training. Training is conducted for up to 10 epochs with early stopping triggered when validation macro-F1 fails to improve for three consecutive epochs; the checkpoint with the highest validation F1 is retained for test evaluation.

4.3 Hardware and Training Time

All experiments are conducted on a single NVIDIA A100 80GB GPU. Training the full BERT-BiLSTM- Attention model on the LIAR dataset requires approximately 4.5 hours per run across 10 epochs with a batch size of 32.

SST-2 training completes in approximately 2 hours due to shorter average sequence length. The HealthClaim-V2 dataset, being smallest, trains in under 1.5 hours. Inference on a held- out test set of 1,000 examples requires approximately 8 seconds at batch size 32, corresponding to a per-example latency of roughly 250 milliseconds including all preprocessing steps.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

4.4 Evaluation Protocol

All datasets are split into training, validation, and held-out test partitions using a stratified random split that preserves the original label proportions in each split. For LIAR and HealthClaim-V2, we use the official train/validation/test splits provided with the datasets to ensure comparability with prior published results. For SST-2, the standard GLUE benchmark split is used. No hyperparameter decisions, including early stopping, are made using test set performance; the test partition is used exclusively for final evaluation.

Reported metrics include accuracy, macro-averaged precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). Macro averaging is preferred over micro averaging for imbalanced datasets because it weights each class equally, avoiding the distortion produced by high performance on majority classes. All metrics are computed over three independent runs with different random seeds, and mean and standard deviation are reported. Confidence intervals at the 95% level are estimated using bootstrap resampling over 1000 resamples of the test set.

V. EXPERIMENTS AND RESULTS

5.1 Datasets

Experiments are conducted on three publicly available benchmark datasets that span distinct prediction tasks and pose different challenges with respect to class balance, sequence length, and domain vocabulary. Table 1 summarises the dataset statistics.

Table 1: Dataset Statistics

Dataset	Task	Train	Val	Test
LIAR	Fake News Detection	10,240	1,284	1,267
SST-2	Sentiment Analysis	67,349	872	1,821
HealthClaim- V2	Medical Misinformation	4,820	602	610

The LIAR dataset contains labelled political statements drawn from PolitiFact, annotated with six-way truthfulness labels that we collapse to a binary real/fake categorisation following prior work. Statements range from one to several sentences in length, and the dataset presents a moderate class imbalance after collapsing. SST-2 provides sentence-level binary sentiment annotations for movie reviews; it is nearly balanced and contains relatively short inputs, making it a useful comparison point for isolating the effect of sequence length on model behaviour. HealthClaim- V2 is a proprietary collection of health-related social media claims annotated for factual accuracy; it presents the strongest class imbalance of the three datasets, with verified misinformation comprising approximately 22% of examples.

5.2 Baseline Models

Five baselines span classical and neural approaches. The logistic regression baseline uses TF-IDF weighted unigram and bigram features with L2 regularisation; the regularisation strength is tuned on the validation set. The BiLSTM-GloVe baseline initialises the embedding layer with 300-dimensional GloVe vectors and fine-tunes them during training; the architecture otherwise mirrors the recurrent component of the proposed model. The TextCNN baseline applies convolutional filters of widths 2, 3, and 4 to GloVe- initialised embeddings and concatenates the resulting max-pooled feature maps before classification. The BERT fine- tuned baseline uses the bert-base-uncased encoder with a single linear classification head over the [CLS] token, trained under identical hyperparameter conditions as the proposed model's encoder. The proposed BERT-BiLSTM-Attention model constitutes the fifth entry and primary contribution.

5.3 Comparative Results

Table 2 presents macro-F1 and accuracy results across all models and datasets. Numbers represent means over three random seeds; standard deviations are below 0.5 points in all cases.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

Table 2: Model Comparison — Macro-F1 (LIAR, HealthClaim) and Accuracy (SST-2)

Model	LIAR F1	SST-2 Acc	Health F1	Av
LR + TF-IDF	54.2	83.1	61.4	5.0
BiLSTM + GloVe	61.7	87.5	67.9	4.0
TextCNN	60.4	88.2	66.3	3.7
BERT (fine-tuned)	67.8	93.4	74.1	2.0
BERT-BiLSTM-Attn (ours)	70.1	94.2	76.6	1.0

The proposed model achieves the best performance across all three datasets. On LIAR, the gain over standalone BERT is 2.3 F1 points, consistent across seeds. On HealthClaim-V2, the gain grows to 2.5 points, a finding we attribute to the minority class constituting a smaller fraction of training data: the BiLSTM and attention mechanism appear to provide a regularisation effect that improves minority-class recall under low-data conditions. On SST-2, the improvement is modest at 0.8 accuracy points, consistent with the expectation that short, stylistically simple sentiment sentences leave less room for sequential modelling to add value.

The gap between TF-IDF logistic regression and all neural models is substantial across datasets, confirming that distributional representations carry significantly more predictive information than sparse frequency-based features for these tasks. The BiLSTM-GloVe and TextCNN baselines are broadly comparable, with TextCNN slightly stronger on SST-2 and BiLSTM-GloVe stronger on the longer-sequence LIAR and HealthClaim datasets, consistent with the theoretical expectation that recurrence offers an advantage as input length increases.

5.4 Ablation Study

Table 3 presents the results of the ablation study, which removes or replaces one component at a time from the full proposed model.

Table 3: Ablation Study Results

Configuration	LIAR F1	Health	SST-2 F1	Acc
Full model	70.1	76.6	94.2	
w/o Attention (mean pooling)	69.2	75.6	93.9	
w/o BiLSTM ([CLS] only)	68.3	74.8	93.5	
w/o Augmentation	68.6	74.5	94.1	
w/o Temperature Scaling	70.0	76.5	94.2	

Removing the attention mechanism and replacing it with mean pooling over BiLSTM outputs reduces LIAR F1 by 0.9 points and HealthClaim F1 by 1.0 points, with negligible effect on SST-2. This pattern suggests that attention's selective aggregation is most beneficial when the input contains heterogeneous content — a mix of relevant and irrelevant propositions — as is typical in multi-sentence claim verification but not in single-sentence sentiment classification.

Replacing the BiLSTM with a feed-forward network applied directly to the [CLS] token representation reduces LIAR F1 by 1.8 points and HealthClaim F1 by 1.8 points, the largest single-component degradation observed. This result provides the strongest evidence that the recurrent head contributes meaningfully beyond what BERT's positional encoding already supplies. The interaction between explicit sequential modelling and transformer-derived contextual features is non-redundant.

DOI: 10.48175/IJARSCT-35622

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

Removing the data augmentation strategy has no measurable effect on SST-2 but reduces LIAR F1 by 1.5 points and HealthClaim F1 by 2.1 points. The larger degradation on HealthClaim is expected given its stronger class imbalance; augmentation compensates for the reduced minority-class signal in this dataset. The absence of augmentation effects on SST-2 confirms that augmentation adds value specifically in imbalanced settings rather than serving as a universal accuracy booster.

Removing temperature scaling has negligible impact on classification accuracy, as expected, since temperature scaling does not change class assignments. Its value lies in calibration quality rather than rank accuracy: ECE increases from below 0.03 to values between 0.07 and

0.11 across datasets when scaling is removed, confirming that the post-training calibration step is essential for downstream use cases that depend on well-formed probability estimates.

5.5 Attention Visualisation

Qualitative inspection of attention weight distributions on held-out examples provides supporting evidence for the mechanism's interpretive value. On LIAR, high-attention tokens cluster around claim-specific terms, temporal markers that imply recency or historical precedent, and source attribution phrases. On HealthClaim-V2, attention concentrates on causal language, medical nomenclature, and quantitative assertions such as percentages and dosage figures. These patterns align with the linguistic theory of how misinformation is constructed, providing a degree of face validity for the model's learned decision boundaries.

VI. DISCUSSION

6.1 Deployment Architecture and Latency Management

Production deployment of transformer-based models demands careful attention to inference latency. The BERT-base encoder alone introduces approximately 60–80 milliseconds of GPU latency per batch of 32 examples on modern accelerator hardware, implying per-example latency in the low single-digit milliseconds for batched asynchronous processing but substantially higher latency for real-time single-example serving. Applications requiring sub-10-millisecond response times — interactive user interfaces, real-time fraud detection, live content moderation — cannot absorb this cost without mitigation.

Knowledge distillation offers the most principled latency reduction strategy. A student model with fewer layers, smaller hidden dimensions, or both is trained to minimise a distillation loss combining the cross-entropy loss on hard labels with a Kullback-Leibler divergence loss against the teacher model's softened output probabilities. The soft probabilities encode the teacher's uncertainty structure and inter-class similarity judgments, providing a richer training signal than hard labels alone. DistilBERT and TinyBERT demonstrate that 6-layer students retain approximately 97% and 99% of teacher performance respectively on GLUE benchmarks at 40–60% of inference cost.

INT8 post-training quantisation reduces memory bandwidth and accelerates matrix multiplications on hardware with INT8 multiply-accumulate units by representing model weights and activations as 8-bit integers rather than 32-bit floats. Dynamic quantisation applies this transformation only to weight matrices, leaving activations in floating point, and can be applied without any retraining using PyTorch's native quantisation API. Static quantisation additionally quantises activations, requiring calibration on a small representative dataset, and yields larger speedups at the cost of slightly higher implementation complexity. Quantisation-aware training, which simulates quantisation noise during fine-tuning, offers the best accuracy-under- quantisation trade-off but requires access to training data and additional compute.

For the specific deployment scenario of asynchronous document batch processing — common in compliance monitoring, news aggregation, and clinical audit — none of the above mitigations may be strictly necessary. A well-designed batching layer that accumulates incoming documents and forwards them in size-64 or size-128 batches can sustain high throughput with standard BERT inference, and the latency perceived by downstream consumers is determined by batch flush interval rather than per-example processing time.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

6.2 Domain Adaptation and Continual Learning

Fine-tuned models experience distribution shift when the statistical properties of production inputs diverge from those of the training corpus. Vocabulary shift is a particularly common cause: specialised jargon, emerging terminology, and evolving entity names all introduce out-of- distribution tokens that the model must handle through subword fallback. Continued pre-training of the BERT encoder on domain-specific unlabelled text — a technique variously called domain-adaptive pre-training or intermediate-task pre-training — is an effective strategy for closing the distribution gap before task-specific fine-tuning. The pre-training objective is identical to the original masked language modelling task; only the corpus changes.

In production environments, labelled data continues to accumulate as human annotators review model predictions, creating an opportunity for iterative model improvement. Naïve periodic retraining on the accumulated dataset is computationally expensive and may discard the regularisation benefits of the original pre-trained initialisation. An experience replay strategy maintains a memory buffer of representative examples from earlier training phases; when the model is updated on new data, a mixed mini-batch drawn from both new and buffered examples is used, preventing catastrophic forgetting of previously learned patterns. Gradient episodic memory and elastic weight consolidation offer more principled alternatives with stronger theoretical guarantees against forgetting, at the cost of additional implementation complexity.

6.3 Fairness, Bias, and Societal Implications

Prediction models trained on naturally collected text inherit the biases present in the annotation process and the underlying corpus. Annotator disagreement on subjective tasks such as veracity assessment is not random: it reflects genuine differences in background knowledge, cultural context, and interpretive norms. Models trained on majority-annotator labels effectively learn the distribution of one demographic slice of the annotator population, potentially systematically underperforming for content produced by or about underrepresented groups.

Demographic parity requires that the proportion of positive predictions be equal across demographic subgroups. Equalised odds requires that both true positive rate and false positive rate be equal across subgroups. These two criteria are algebraically incompatible in general when group base rates differ, forcing practitioners to choose which fairness notion is most appropriate for their application context. For high-stakes prediction tasks — decisions about clinical treatment, legal status, or financial access — this choice is ethically significant and should involve domain experts and affected communities rather than being determined unilaterally by the modelling team.

Adversarial de-biasing introduces a secondary adversary network trained to predict a protected attribute from the model's intermediate representations; the primary model is then trained to simultaneously maximise task performance and minimise the adversary's accuracy, encouraging representations that are uninformative about the protected attribute. Re-weighting training examples to balance demographic representation is a simpler alternative. Both strategies reduce bias along the targeted dimension but may introduce trade-offs with aggregate accuracy or with fairness along unmeasured dimensions, highlighting the importance of comprehensive auditing before deployment.

6.4 Interpretability and User Trust

Real-world deployment of prediction systems in consequential domains increasingly faces regulatory and organisational requirements for explainability. The European Union's General Data Protection Regulation establishes a right to explanation for automated decisions affecting individuals, and the United States Food and Drug Administration has issued guidance documents on transparency requirements for clinical AI tools. Meeting these requirements demands interpretability mechanisms that are both technically rigorous and communicable to non- expert stakeholders.

LIME (Local Interpretable Model-Agnostic Explanations) generates explanations by perturbing the input around a specific example and fitting a locally linear surrogate model to the prediction surface. SHAP (SHapley Additive exPlanations) allocates prediction credit to input features using a game-theoretic framework that satisfies a set of axiomatic desiderata including local accuracy, consistency, and missingness. Integrated Gradients computes feature

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

attributions by integrating the gradient of the output with respect to each input token along a path from a baseline input to the actual input, satisfying the sensitivity axiom that LIME violates.

For the specific architecture proposed in this paper, attention weights offer a fast and easily visualised approximation to feature importance. Their limitations are well-documented: attention is not equivalent to gradient- based attribution, and high attention weight on a token does not guarantee that perturbing that token would meaningfully affect the prediction. Responsible deployment therefore surfaces attention weights as a preliminary exploration tool while directing users who require formal attribution to gradient-based methods. Combining both signals, using attention for fast initial triage and gradient attribution for detailed case review, balances interpretability cost against depth of explanation.

6.5 Limitations

Several limitations of the proposed system merit explicit acknowledgment. The experimental evaluation is restricted to English-language datasets; the architectural approach is language-agnostic in principle, but the choice of BERT-base-uncased as the backbone introduces English- specific tokenisation assumptions that would need to be addressed through multilingual or language-specific pre- trained models for non-English deployments. The datasets used, while publicly available and widely cited, represent a narrow slice of the full range of prediction tasks to which the system might be applied; practitioners should conduct their own domain-specific evaluation before drawing conclusions about expected production performance.

The hybrid architecture introduces additional hyperparameters relative to a simple BERT fine-tuning approach: the BiLSTM hidden size, number of layers, inter- layer dropout, and attention projection dimension must all be set. While the configurations reported in this paper generalise reasonably well across the three benchmarks tested, there is no guarantee that they are optimal for a new domain, and a validation set hyperparameter sweep should be considered standard practice for new deployments.

VII. CONCLUSION

This paper has presented a comprehensive AI- powered text prediction system that integrates the contextual encoding strength of BERT with the sequential modelling capacity of a bidirectional LSTM, unified through an additive attention mechanism and completed by a post-training probability calibration stage. The system addresses the full prediction workflow from raw text acquisition through preprocessing, augmentation, deep encoding, sequential classification, and calibration, with each stage designed for independent replacement to accommodate evolving model capabilities and deployment constraints.

Experimental results across three benchmark datasets confirm that the hybrid BERT-BiLSTM-Attention architecture consistently outperforms standalone baselines, with the greatest gains on tasks characterised by multi- sentence inputs, class imbalance, or complex argumentative structure. The ablation study demonstrates that each proposed component — the recurrent head, the attention mechanism, the augmentation strategy, and the calibration step — contributes meaningfully to the overall system, and that their contributions are largely non-overlapping.

The discussion of deployment considerations identifies knowledge distillation and quantisation as effective strategies for latency reduction, domain-adaptive pre-training and experience replay for handling distribution shift, adversarial de-biasing and example re-weighting for fairness improvement, and integrated gradients alongside attention visualisation for interpretability support. These considerations collectively constitute a roadmap for moving from a research prototype to a production system that meets the operational, ethical, and regulatory requirements of consequential deployment contexts.

Future work will explore four directions. First, multimodal fusion that combines text representations with structured numerical features through a cross-attention bridge layer could extend the system's applicability to domains where both data types are available. Second, a systematic comparison of lightweight encoder variants — DistilBERT, ALBERT-lite, and MobileBERT — against the bert-base backbone would clarify the accuracy-latency trade-off curve in greater

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

resolution than is possible from published benchmarks alone. Third, an evaluation of the system under realistic distribution shift conditions, where training and deployment corpora differ in provenance, time period, and authorial style, would provide more operationally relevant performance estimates than held-out test sets from the same distribution. Fourth, a user study examining how practitioners interpret and act upon attention-based and gradient-based explanations in domain-specific decision-making contexts would ground the interpretability claims of this work in human behavioural evidence.

The intersection of natural language understanding and predictive modelling remains one of the most productive areas in applied machine learning. The architecture and pipeline presented here offer a principled and extensible foundation that connects research advances to production requirements, and the experimental evidence confirms that a thoughtfully combined hybrid design outperforms its individual components across a range of realistic prediction setting.

REFERENCES

[1] Kohad, Reshma, Nidhi Khare, Sachin Kadam, Nidhi, Vishal Borate, and Yogesh Mali. "A Novel Approach for Identification of Information Defamation Using Sarcasm Features." In the International Conference on Information Technology and Intelligence, pp. 159-170. Singapore: Springer Nature Singapore, 2024.

[2] Mali, Y. (2009). Identification of New Protein-protein Interaction of the Amyotrophic Lateral Sclerosis-linked Mutant G93A Human Superoxide Dismutase and Its Functional Implication. Tel Aviv University.

[3] Mali, Yogesh, and Viresh Chapte. “Grid based authentication system.” International Journal 2, no. 10 (2014).

[4] Lokre, Amit, Sangram Thorat, Pranali Patil, Chetan Gadekar, and Yogesh Mali. "Fake image and document detection using machine learning." International Journal of Scientific Research in Science and Technology (IJSRST) 5, no. 8 (2020): 104-109

[5] Y. K. Mali, S. A. Darekar, S. Sopal, M. Kale, V. Kshatriya and A. Palaskar, "Fault Detection of Underwater Cables by Using Robotic Operating System," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 1-6, doi: 10.1109/ICCST59048.2023.10474270.

[6] Y. K. Mali, S. Dargad, A. Dixit, N. Tiwari, S. Narkhede and A. Chaudhari, "The Utilization of Block-chain Innovation to Confirm KYC Records," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 1-5, doi: 10.1109/ICCST59048.2023.10530513.

[7] Lonari, P., Jagdale, S., Khandre, S., Takale, P., & Mali, Y. (2021). Crime awareness and registration system. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 8(3), 287-298.

[8] Asreddy, R., Shingade, A., Vyavhare, N., Rokde, A., & Mali, Y. (2019). A survey on secured data transmission using RSA algorithm and steganography. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 4(8), 159-162.

[9] Pathak, J., Sakore, N., Kapare, R., Kulkarni, A., & Mali, Y. (2019). Mobile rescue robot. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 4(8), 10-12.

[10] Yogesh Mali and Tejal Upadhyay, “Fraud Detection in Online Content Mining Relies on the Random Forest Algorithm,” SWB, vol. 1, no. 3, pp. 13–20, Jul. 2023, doi: 10.61925/SWB.2023.1302.

[11] Chougule, Shivani, Shubham Bhosale, Vrushali Borle, and Vaishnavi Chaugule. "Prof. Yogesh Mali,“Emotion Recognition Based Personal Entertainment Robot Using ML & IP." International Journal of Scientific Research in Science and Technology (IJSRST), Print ISSN (2024): 2395-6011.

[12] Mali, Y. (2023). TejalUpadhyay, “. Fraud Detection in Online Content Mining Relies on the Random Forest Algorithm ”, SWB, 1 (3), 13–20.

[13] Modi, S., Mane, S., Mahadik, S., Kadam, R., Jambhale, R., Mahadik, S., & Mali, Y. (2024). Automated Attendance Monitoring System for Cattle through CCTV. REDVET-Revista electrónica de Veterinaria, 25(1), 2024.

[14] N. Nadaf, G. Chendke, D. S. Thosar, R. D. Thosar, A. Chaudhari and Y. K. Mali, "Development and Evaluation of RF MEMS Switch Utilizing Bimorph Actuator Technology for Enhanced Ohmic Performance," 2024 International

Conference on Control, Computing, Communication and Materials (ICCCCM), Prayagraj, India, 2024, pp. 372-375, doi: 10.1109/ICCCCM61016.2024.11039926.
[15] D. Das et al., "Antibiotic susceptibility profiling of Pseudomonas aeruginosa in nosocomial infection," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10723982.
[16] Y. K. Mali, L. Sharma, K. Mahajan, F. Kazi, P. Kar and A. Bhogle, "Application of CNN Algorithm on X-Ray Images in COVID-19 Disease Prediction," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 1-6, doi: 10.1109/ICCST59048.2023.10726852.
[17] Inamdar, Faizan, Dev Ojha, C. J. Ojha, and D. Y. Mali "Job Title Predictor System." International Journal ofAdvanced Research in Science, Communication and Technology (2024): 457–463.
[18] Jagdale, Sudarshan, Piyush Takale, Pranav Lonari Shraddha, Khandre, and Yogesh Mali. "Crime Awareness and Registration System." International Journal of Scientific Research in Science and Technology 5, no. 8 (2020).
[19] Mali, Yogesh. "NilaySawant,"Smart Helmet for Coal Mining,"." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 3.
[20] P. Koli, V. Ingale, S. Sonavane, A. Chaudhari, Y. K. Mali and S. Ranpise, "IoT-Based Crop Recommendation Using Deep Learning," 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), Prayagraj, India, 2024, pp. 391-395, doi: 10.1109/ICCCCM61016.2024.11039888.
[21] Mali, Yogesh Kisan. "Marathi sign language recognition methodology using Canny's edge detection." Sādhanā 50, no. 4 (2025): 268.
[22] Y. K. Mali and A. Mohanpurkar, "Advanced pin entry method by resisting shoulder surfing attacks," 2015 International Conference on Information Processing (ICIP), Pune, India, 2015, pp. 37-42, doi: 10.1109/INFOP.2015.7489347.
[23] Hajare, R., Hodage, R., Wangwad, O., Mali, Y., & Bagwan, F. (2021). Data security in cloud. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 8(3), 240-245.
[24] Dhote, D., Rai, P., Deshmukh, S., & Jaiswal, A. Prof. Yogesh Mali," A Survey: Analysis and Estimation of Share Market Scenario. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN, 2456-3307.
[25] T. S. Ruprah, V. S. Kore and Y. K. Mali, "Secure data transfer in android using elliptical curve cryptography," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-4, doi: 10.1109/ICAMMAET.2017.8186639.
[26] Bhongade, A., Dargad, S., Dixit, A., Mali, Y.K., Kumari, B., Shende, A. (2024). Cyber Threats in Social Metaverse and Mitigation Techniques. In: Somani, A.K., Mundra, A., Gupta, R.K., Bhattacharya, S., Mazumdar, A.P. (eds) Smart Systems: Innovations in Computing. SSIC 2023. Smart Innovation, Systems and Technologies, vol 392. Springer, Singapore. https://doi.org/10.1007/978-981-97-3690-4_34.
[27] A. More, S. Khane, D. Jadhav, H. Sahoo and Y. K. Mali, "Auto-shield: Iot based OBD Application for Car Health Monitoring," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-10, doi: 10.1109/ICCCNT61001.2024.10726186.
[28] A. More, O. L. Ramishte, S. K. Shaikh, S. Shinde and Y. K. Mali, "Chain-Checkmate: Chess game using blockchain," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725572.
[29] P. Shimpi, B. Balinge, T. Golait, S. Parthasarathi, C. J. Arunima and Y. Mali, "Job Crafter - The One-Stop Placement Portal," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-8, doi: 10.1109/ICCCNT61001.2024.10725010.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

[30] A. Chaudhari et al., "Cyber Security Challenges in Social Meta-verse and Mitigation Techniques," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-7, doi: 10.1109/MITADTSoCiCon60330.2024.10575295.
[31] Kale, H., Aswar, K., Yadav, Y.M.K. and Mali, D.Y., 2024. Attendance marking using face detection. International Journal of Advanced Research in Science, Communication and Technology, 417424.
[32] Mali, Yogesh Kisan, Vijay Rathod, Sweta Dargad, and Jyoti Yogesh Deshmukh. "Leveraging Web 3.0 to Develop Play-to-Earn Apps in Healthcare using Blockchain." In Computational Intelligence and Blockchain in Biomedical and Health Informatics, pp. 243-257. CRC Press, 2024.
[33] Koli, Pooja, Vinod Ingale, Sonali Sonavane, Ashvini Chaudhari, Yogesh Kisan Mali, and Shivam Ranpise "IoT-BasedCrop Recommendation Using Deep Learning." In 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), pp. 391 - 395. IEEE, 2024.
[34] Kulkarni, Varsha G., Vishal Borate, and Yogesh Mali. "An Intelligent Ventilation Bag Featuring Automated Pressure Control and Variable Oxygen Range.," International Journal of Advanced Research in Science, Communication and Technology, Volume- 6, Issus- 2, pp: 238–255, 2026.
[35] Pisote, Anita, Yogesh Mali, and Vishal Borate. "An AI-Driven Framework for Digitized Audiological Reporting Based on Audiogram Analysis." International Journal of Advanced Research in Science, Communication and Technology, Volume- 6, Issus- 2, pp: 256–270, 2026.
[36] Lilhare, Shweta G., Vishal Borate, and Yogesh Mali. "An AI & ML based BM25-Driven Methodology for Shortlisting Job Applicant Resumes." International Journal of Advanced Research in Science, Communication and Technology, Volume- 6, Issus- 2, pp: 224–237, 2026.
[37] Mahajan, Krishnal, Sumant Bhange, Prajakta Gade, and Yogesh Mali "Guardian Shield: Real Time Transaction Security."
[38] Bhoye, Tejaswini, Aishwarya Mane, Vandana Navale, Sangeeta Mohapatra, Sandeep Chitalkar, Vishal Borate, and Yogesh Mali. "A role of machine learning algorithms for demand based Netflix recommendation system." In Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025), vol. 2, pp. 212-220. 2025.
[39] Umar Mulani, Dr Vinod Ingale, Rais Mulla, Ankita Avthankar, Yogesh Mali, and Vishal Borate. "Optimizing Pest Classification in Oil Palm Agriculture using Fine-Tuned GoogleNet Deep Learning Models." (2025).
[40] Yogesh, Jyoti. "and Classification for Varicose Veins." Data-Centric Artificial Intelligence for Multidisciplinary Applications (2024): 114.
[41] Y. K. Mali, V. U. Rathod, N. P. Sable, R. R. Rathod, N. A. Rathod and M. N. Rathod, "A Technique for Maintaining Attribute-based Privacy Implementing Blockchain and Machine Learning," 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India, 2023, pp. 1-4, doi: 10.1109/GCITC60406.2023.10426183.
[42] Mali, Y.K., Rathod, V.U., Borate, V.K., Chaudhari, A., Waykole, T. (2024). Enhanced Pin Entry Mechanism for ATM Machine by Defending Shoulder Surfing Attacks. In: Roy, N.R., Tanwar, S., Batra, U. (eds) Cyber Security and Digital Forensics. REDCYSEC 2023. Lecture Notes in Networks and Systems, vol 896. Springer, Singapore. https://doi.org/10.1007/978-981-99-9811-1_41.
[43] Mali, Yogesh. "TejalUpadhyay,"." Fraud Detection in Online Content Mining Relies on the Random Forest Algorithm", SWB 1, no. 3 (2023): 13-20.
[44] Mali, Y., Rathod, V. U., Kulkarni, M. M., Mokal, P., Patil, S., Dhamdhere, V., & Birari, D. R. (2023). A comparative analysis of machine learning models for soil health prediction and crop selection. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 811-828.
[45] Y. K. Mali, V. U. Rathod, M. D. Salunke, S. B. Satish, P. Dhamdhere and R. R. Rathod, "Role of IoT in Coal Miner Safety Helmets," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg, Germany, 2023, pp. 221-225, doi: 10.1109/ICCCMLA58983.2023.10346793.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

[46] Y. Mali, V. U. Rathod, R. S. Tambe, R. Shirbhate, D. Ajalkar and P. Sathawane, "Group-Based Framework for Large Files Downloading," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-4, doi: 10.1109/ICCCNT56998.2023.10308339.
[47] Y. K. Mali, V. U. Rathod, S. Dargad, V. N. Kapure, N. Vyawahare and S. Gajarlewar, "Development of ROS(Robotic Operating System) for Fault Detection of Underwater Cables," 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 2023, pp. 956-961, doi: 10.1109/R10-HTC57504.2023.10461858.
[48] Mali, Y.K., Rathod, V.U., Mali, N.D., Mahajan, H.C., Nandgave, S., Ingale, S. (2025). Role of Block-Chain in Medical Health Applications with the Help of Block-Chain Sharding. In: Madureira, A.M., Abraham, A., Bajaj, A., Kahraman, C. (eds) Hybrid Intelligent Systems. HIS 2023. Lecture Notes in Networks and Systems, vol 1227. Springer, Cham. https://doi.org/10.1007/978-3-031-78931-1_8.