AISA-BS: A Multimodal Employment Screening Framework Integrating Transformer-Based Semantic Analysis and IoT-Driven Behavioral Sensing

Wenzhao Zhang
Zhengzhou University of Science and Technology, Zhengzhou 450064, Henan, China
E-mail: wenzhao-zhang@outlook.com

Keywords: Multimodal Employment Screening, Transformer-based NLP, IoT Behavioral Sensing, BiLSTM Temporal Analysis, Tensor Fusion, Human-Centric AI, DAiSEE Dataset

Received: May 30, 2025

With the rise of digital hiring platforms, it has become increasingly challenging to evaluate job candidates using only resumes and interviews accurately. Traditional screening methods often overlook important behavioral and contextual cues, which can lead to poor hiring decisions. To overcome these limitations, there is a growing need for more comprehensive screening systems. This paper proposes AISA-BS (Artificial Intelligence Semantic Analysis and Behavioral Sensing), a multimodal employment screening framework designed to improve candidate evaluation by combining language analysis with behavioral data collected through IoT devices. AISA-BS leverages Transformer-based NLP models (e.g., BERT) to analyze unstructured text inputs like resumes and interview transcripts. It uses IoT-enabled sensors to capture behavioral data—such as gaze, posture, and stress—from candidates in simulated job environments. These multimodal signals are fused through tensor decomposition and cross-modal attention, and interpreted using a BiLSTM-based behavioral engine for temporal analysis. Experiments were conducted using the DAiSEE dataset, which includes video-based affective state annotations. The proposed model achieved a classification F1-score of 93.2%, reducing the Mean Absolute Error (MAE) to 3.1%, outperforming BERT+MLP and MM-DNN baselines. In conclusion, AISA-BS sets a new benchmark for intelligent, fair, and context-aware employment screening by combining deep semantic insight with behavioral interpretation.

Povzetek: Članek obravnava avtomatizirano zaposlovanje v povezavi z vedenjskimi in kontekstnimi signali kandidatov. Predlaga multimodalni sistem AISA-BS, ki združuje semantično analizo (BERT) in IoT-vedenjsko zaznavanje, povezano s tensorno fuzijo in BiLSTM-temporalno analizo.

1 Introduction

Rapid digital change throughout all areas of the economy puts companies under increased pressure to enhance the accuracy, impartiality, and responsiveness of their decision-making procedures regarding employment. This shift is largely driven by the digital revolution, which has accelerated the need for more accurate, impartial, and responsive employment decision-making [1]. The traditional methods of hiring, which are usually focused on static resumes, rigorous interviews, and subjective assessments, cannot simultaneously reflect the various characteristics of human potential and workplace compatibility [2]. These methods generally depend on static resumes and formal interviews alone, which restrain them from being able to capture the complexity of candidate behavior. These approaches not only require time and are susceptible to cognitive biases, but they also neglect important behavioral and contextual factors affecting job performance [3]. This restriction greatly reduces the effectiveness of conventional hiring schemes. Given the changing workforce dynamics in line with the industry 5.0 model, which emphasizes human-AI cooperation, personalization, and resilience, there is a pressing need for smart, data-driven systems capable of thoroughly assessing applicants and matching them with job roles [4].

1.1 Problem findings

In-depth analysis and business processes have uncovered core weaknesses in traditional hiring processes, notably the inability to incorporate real-time behavioral evaluation. This weakness prevents the determination of how job applicants react to working pressures, learn through fluid situations, and express interest in mock working conditions. Lack of such behavior patterns is a core weakness, since such traits are critically relevant to real job performance and situational adaptability judgments. Second, although sometimes solely for keyword-based filtering or sentiment analysis, artificial intelligence has been widely employed in hiring. Although used extensively, AI is generally restricted to simpler uses such as keyword matching or tone checking, not grasping more sophisticated meaning. However, existing AI-powered recruitment tools do not capture deeper semantic patterns that reflect soft skills, intrinsic motivation, or ethical

alignment — issues of the highest significance to real-world job performance. This outcome was brought about by circumstances [6]. Thirdly, dynamic feedback may change decision-making to fit evolving organizational requirements and experience. These models lack dynamic input, limiting their ability to adapt to evolving organizational needs.

1.2 Proposed solution and methodology

By proposing a new job decision-making model, AISA-BS (Artificial Intelligence Semantic Analysis and Behavioral Sensing), this study seeks to overcome these constraints. This method combines two significant fields of technology: (1) behavioral sensing based on the Internet of Things, and (2) semantic analysis driven by artificial intelligence. The methodology consists of multiple interconnected stages that work together to facilitate context-sensitive and dynamic decision-making during recruitment [7]. The method is split into several distinct stages. Initially, methods from the domain of Natural Language Processing (NLP) are used to draw semantic information from unstructured text input. This is done across the first phases of the process. Among the many characteristics in this category are internet profiles, interview records, and CVs [8]. These methods allow one to identify linguistic signals linked to cognitive, emotional tone, and communicative style in line with the results. Potential workers are engaged in activities simulating their employment during the same time frame, and IoT-enabled sensors track them. These sensors can collect a range of behavioral data, including micro-interactions, levels of engagement, and stress reactions. Combining these multimodal data streams with a multi-layered decision engine using machine learning models allows one to evaluate whether or not applicants fit the positions offered. Adaptive learning allows the system to constantly change its decision-making patterns by considering comments and past performance of the system. The system's capacity to alter these patterns creates the realization of this potential. All things considered; this leads to dynamic, context-aware, objective ideas.

1.3 Research objectives

This research project has several main objectives that it wants to achieve, some of which are mentioned below:

- To compare AISA-BS performance against traditional employment screening techniques in delivering ≥5% F1-score improvement and lower MAE and RMSE values for different sample sizes.
- To construct and cross-validate an adaptive learning model, measured in terms of how it can update classification results based on task information in real-time.
- To cross-validate fairness and explainability by comparing subgroup performance (e.g., age, gender) and via interpretability metrics such as attention attribution heatmaps.
-

2 Literature survey

The academic community has recently been very interested in using artificial intelligence (AI) and the Internet of Things (IoT) to manage and hire personnel. This has led to a lot of attention. Scientists have put more emphasis on AI and IoT in hiring as the urgent demand for more efficient and context-focused hiring solutions has grown. Researchers have stopped using traditional rule-based systems in the last few years. This change has helped intelligent models that use deep semantic analysis and real-time psychological data become more popular. This shift taps into the expanding corporate need for smarter, more equitable, and more contextual styles of hiring that can accommodate changing workplace conditions. This section introduces the recent developments in AI and IoT-based recruitment frameworks to place the development and contribution of the AISA-BS model into perspective. This goal will be reached by pointing out the most important strengths and areas for improvement in the fields necessary for the model to grow.

2.1 AI in employment decision-making

Artificial intelligence has also had a significant impact on pre-screening and candidate evaluation, as it has transformed the hiring process. Both regions have witnessed unprecedented limitations on initial deployments. These limitations have been extensively documented across different systems. Black and van Esch [9] found that the first systems used rule-based filtering and keyword-matching algorithms. Both algorithms often led to bias or the lack of a significant feature in the context. The study was done to learn how well these early systems worked. Chakraborty et al. [10] found that natural language processing (NLP) and deep learning are the newest methods for analyzing profiles on social media, interview transcripts, and resumes online. These techniques have been shown to facilitate improved interpretation of various textual information. The algorithms can detect cognitive and affective characteristics, recognize linguistic habits, and analyze communication styles. This helps the system to build a more detailed and complex representation of candidate profiles, making possible more precise recruitment choices.

Even if there have been significant improvements, using technology powered by artificial intelligence remains a challenge. People can now use these technologies. Liem et al. [11] say that most models focus on text data and don't use real-time behavioral signals in their methods. This is the case because most models don't have this feature. Because of this, these people are less able to judge prospects and make good decisions in a wide range of situations. To prove this point, most models only focus on text data. Research by Sanchez-Monedero et al., [12] found that black-box algorithms could also raise concerns about openness. Because of this, there is a chance that there will be more problems with justice, openness, and following labor laws.

2.2 Semantic analysis in recruitment

Natural Language Processing (NLP) methods enable companies to transition from simple keyword matching to semantic text understanding, representing complex linguistic and contextual patterns. This lets them increase their capacity beyond the straightforward keyword identification. Candidate communications are examined using sentiment analysis, topic modeling, and named entity identification to assess the candidate's soft skills, intentions, and emotional tone [13]. This lets one assess the emotional tone of the candidate. Advanced models like BERT and GPT-based systems have improved the ability to identify contextual meanings in job applications, reducing the need for manual screening. This development represents a long-term shift toward reducing manual screening activity using automated and intelligent systems. The time spent on manual screening and the time spent on it have both fallen.

Furthermore, whereas semantic analysis greatly enhances knowledge of candidate profiles, its efficacy is constrained without corroborating data. Omar et al. [14] and colleagues indicate that the most successful way to assess candidates includes interactive and behavioral environments. Such environments, for instance, include stress-handling simulations or role-play situations. Should text interpretation be seen in isolation, applicants may be misled.

2.3 IoT-Based behavioral sensing in human-centric systems

Amongst many others, human-centric systems' Internet of Things (IoT) applications have been proven effective in various sectors, including healthcare, smart learning, and safety monitoring. These applications have also been effective in supporting the human-centric systems through facilitating improved decision-making and optimization. Bhattacharyya et al. [15] claim that sensors integrated into smart environments can gather real-time data on motion, gaze, stress, heart rate, and other physiological and behavioral markers connected with the surroundings. Sensors included in smart environments have this capacity. Using these signals is helpful, particularly in creating behavioral profiles of people in dynamic, task-driven settings.

Although it's innovative, employing the Internet of Things in hiring does offer a lot of potential, particularly for enhancing data analysis and monitoring behavior in real time. Most available studies would rather concentrate on staff monitoring than applicant assessment. The study does not primarily focus on candidate assessment. Sanchez-Iborra et al. [16] claim the difficulty lies in using Internet of Things technology to assess candidates in simulations mimicking job settings. This assessment must be conducted without violating the privacy of possible applicants or creating new forms of prejudice. A key concern remains the ethical use of this technology in decision-making contexts.

2.4 Hybrid decision models and adaptive hiring frameworks

Hybrid decision models are gaining increasing attention as they facilitate accurate talent assessments. These models combine various AI methods—like natural language processing, behavior detection, and predictive analytics—to accomplish recruitment objectives. Chander, B. [17], it is suggested that these models usually include features obtained by natural language processing, structured candidate profiles, and historical performance data. Zhang, J., & Tao, D. [18], in these models are usually used in recruitment systems, as different studies have reported. Conversely, a few studies use real-time behavioral data; even fewer apply adaptive learning techniques to improve decision logic over long durations. Having this limitation is quite significant. The AISA-BS model tries to close this gap by incorporating Internet of Things-based behavioural sensing techniques with artificial intelligence-controlled semantic analysis. This is made possible by applying a single decision framework to solve the problem. Unlike conventional systems, which do not permit ongoing learning and dynamic adaptation, it handles the issues connected with context sensitivity, fairness, and openness. Traditional systems do not operate this way. To contextualize AISA-BS's contributions more effectively, Table 1 summarizes and contrasts the salient features, performance evaluation criteria, and limitations of state-of-the-art intelligent employment screening models of recent times. Table 2 Summarizes the research gap.

Table 1: Comparison of State-of-the-Art Methods vs. AISA-BS

Model	Features	F1-Score	MAE (%)	RMSE (%)	Key Limitations
BERT + MLP	Text-only analysis using Transformer embeddings and MLP classification	86.7%	6.2	8.3	No behavioral data; lacks context awareness; non-adaptive; black-box predictions
BiLSTM + MLP	Temporal modeling of text using BiLSTM and MLP	88.4%	6.7	8.7	Limited to semantic sequences; no multimodal signals; lacks behavioral modeling
MM-DNN	Early fusion of multimodal features using dense layers	91.1%	4.9	6.9	Simple fusion strategy; lacks cross-modal attention; non-adaptive
AISA-BS	Transformer NLP + BiLSTM for behavior + Tensor-based fusion with cross-modal attention	93.2%	3.1	4.8	-

Table 2: Summary of Research Gaps and AISA-BS Contributions

Identified Research Gap	Limitations in Existing Work	How AISA-BS Addresses It
Limited semantic depth in AI-based recruitment	Many systems rely on simple keyword matching or sentiment detection only (Black & van Esch, 2020)	Uses fine-tuned BERT to capture contextual, cognitive, and emotional nuances in text.
Lack of behavioral sensing integration	Prior models ignore or minimally use real-time behavioral traits (Liem et al., 2018)	Employs IoT-enabled sensors to capture behavioral cues like gaze, posture, and stress
Poor multimodal data fusion	Existing systems use early or naïve fusion methods (e.g., MM-DNN)	Applies Tucker decomposition and cross-modal attention for deep feature fusion
No adaptive learning or context sensitivity	Traditional frameworks lack a dynamic response to feedback and evolving patterns.	Implements a modular, feedback-driven learning loop to adjust over time

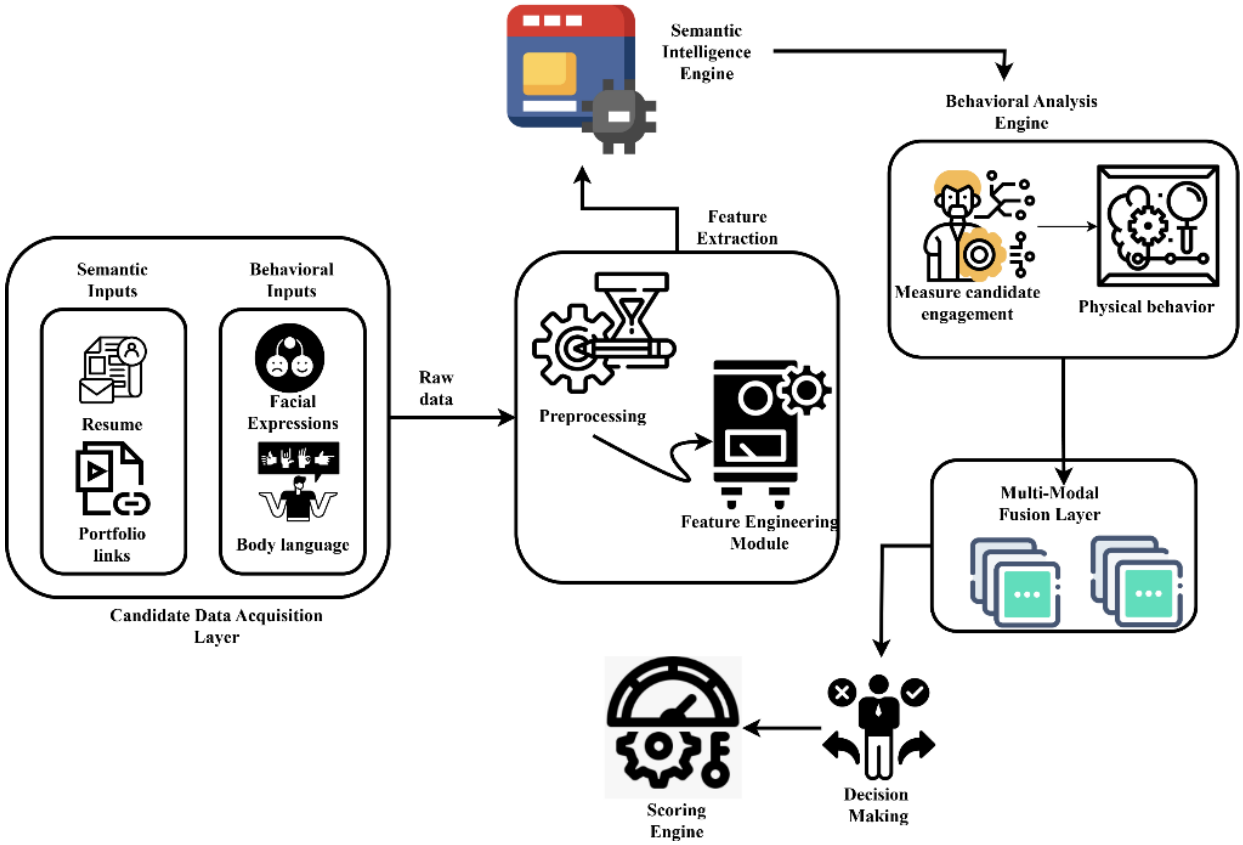

Figure1. Overall proposed workflow of AISA-BS (Artificial Intelligence Semantic Analysis with Behavioral Sensing) Framework

3 Proposed framework AISA-BS (artificial intelligence semantic analysis with behavioral sensing)

Figure 1 shows the proposed system, AISA-BS, which is short for Artificial Intelligence Semantic Analysis with Behavioral Sensing. This is among the systems now being considered. We have made this mechanism accessible to you. Specifically, it is a hybrid intelligent decision-making model meant to transform the process of screening and choosing candidates for employment. Specifically, it is a hybrid intelligent decision-making model meant to transform the process of screening and selecting candidates for employment. This model introduces a context-aware and fair mechanism tailored for modern hiring demands. This project aims to offer a more efficient means of hiring. Semantic analysis, driven by artificial intelligence, and behavioral sensing, based on information technology, are brought together under the canopy of a unified adaptive framework. The platform combines AI-enabled semantic analysis and IoT-enabled behavioral sensing—two complementary science fields—into one framework that provides an optimized assessment of candidates. This effort is being done to enable fair employment decisions that consider the situation's particulars and are based on the facts. To obtain insightful

analysis, we will look at the candidates' spoken material and their conduct under simulated job responsibilities. The aim is to better understand the people seeking the post. Given that this is the case, it will be feasible to achieve the objectives effectively. Once its implementation is finished, the AISA-BS model will move through three key phases: the first is data collection, the second is feature extraction and fusion, and the third is the creation of intelligent judgments. The whole procedure depends on every one of these stages. The technique relies on each of these steps. To effectively finish the process, every stage must be completed. These stages are intended to be modular so that ongoing changes in response to feedback from recruiters, behavioral trends, and changing position needs may be allowed. This modular architecture supports ongoing system refinement according to recruiter comments and changing behavioral patterns.

The AISA-BS method has made a big difference in the field of smart hiring decisions. This progress has been made possible by the technique put in place. The methodology improves the traditional hiring method by providing a complete and data-driven way to decide if a candidate is right for the job. This is done using advanced semantic analysis and behavioral sensing methods. The model can capture both cognitive and emotional states that affect decision-making. Still, it uses a mix of different data types, such as spoken language and physiological signs. This lets the model see both kinds of states at the same time.

When you compare the AISA-BS model to existing models used to make hiring decisions, you can see significant improvements in how accurate, easy to understand, and adaptable the model is expected to be. The many experimental tests that were done showed that these benefits are real. Semantic analysis can correctly understand the subtle language of applicants by gathering information about the context from their written or spoken answers. We can do this using a deep learning architecture based on the Transformer. In contrast, behavioral sensing uses sensors connected to the Internet of Things (IoT) to accurately track physical and mental signs, like stress levels, attention levels, and emotional responses, while the interview is going on. This ensures the integrity and reliability of the interview evaluation process.

3.1 Candidate data acquisition layer (advanced signal modeling)

Candidate Data Acquisition Layer, described as follows: Sensor signal acquisition using the Fourier-Stochastic Composite Model:

$$x_i(t) = \sum_{k=1}^{K} A_{ik} \cdot \sin(2\pi f_{ik} t + \phi_{ik}) + \eta_i(t), \quad \eta_i(t) \sim \mathcal{N}(0, \sigma_i^2) \tag{1}$$

Where in Equation (1),

$i \in \{1,2,\ldots,N\}$ denotes sensor index

K : Number of signal components per sensor

$\eta_i(t)$: Additive white Gaussian noise modeled as $\mathcal{N}(0, \sigma_i^2)$.

σ_i^2 denotes the variance of noise specific to the i^{th} sensor.

The full input tensor is $\mathcal{X} \in \mathbb{R}^{N \times T}$

The data acquisition phase in the AISA-BS system involves collecting a fusion of semantic and behavioral signals using both natural language interfaces and LoT-based sensors. Each behavioral channel, such as heart rate variability, eye-tracking, or galvanic skin response, is modeled as a superposition of sinusoidal basis functions perturbed by Gaussian noise, formally expressed as $x_i(t) = \sum_{k-1}^{K} A_{ik} \cdot \sin(2\pi f_{ik} t + \phi_{ik}) + \eta_i(t)$, where $\eta_i(t) \sim \mathcal{N}(0, \sigma_i^2)$ represents stochastic fluctuations. Aggregating across multiple sensors results in a multidimensional input tensor $\mathcal{X} \in \mathbb{R}^{N \times T}$, where N denotes the number of sensors and T the time series length. This formulation captures temporal dynamics and inter-sensor dependencies essential for modeling complex human behavior during candidate evaluation.

3.2 Preprocessing and feature engineering (contextual embedding and normalization)

Preprocessing and Feature Engineering is Generalized contextual embedding using a layer-weighted Transformer ensemble: $\mathrm{E} = \sum_{l=1}^{L} \alpha_l \cdot \mathrm{h_l}$.

where, α_l is the attention weight of layer l and h_l is the l-th Transformer hidden representation. Contextual embedding E is a weighted sum over all layers L, with scalar weights αl learned during the training time and such that $\sum_{l=1}^{L} \alpha_l = 1$.

For sensor signal preprocessing using multivariate Z-normalization:

$$\mathrm{x}_i' = \Sigma^{-1/2}(\mathrm{x}_i - \mu) \tag{2}$$

$\mathrm{x}_i \in \mathbb{R}^d$: Feature vector

μ : Mean vector

Σ : Covariance matrix of all samples

The transformed vector x_i' represents the normalized version of x_i, obtained via multivariate Z-normalization.

In Equation (2) once raw data is collected, semantic and behavioral inputs undergo preprocessing using domain-specific statistical transformations. Semantic inputs, such as candidate responses, are processed using Transformer-based encoders like BERT, wherein the contextual embedding for a token i is aggregated across layers as $\mathrm{E}_i \sum_{=l-1}^{L} \alpha_l \cdot \mathrm{h}_i^l$, with α_l being learned scalar weights satisfying $\sum_{l=1}^{L} \alpha_l = 1$. Simultaneously, behavioral vectors are normalized using a multivariate z-score transformation: $\mathrm{x}_i' - \Sigma^{-1/2}(\mathrm{x}_i - \mu)$, where Σ is the covariance matrix and μ is the sample mean vector. This ensures uniform scaling and removes bias across modalities, preparing the dataset for robust feature extraction.

3.3 Semantic intelligence engine (multi-head attention and positional encoding)

Full multi-head attention with residual and normalization is given by in Equation (3),

$$\mathrm{MHAtt(X)} = \mathrm{LayerNorm}\left(\mathrm{X} + Concat_{i-1}^{H}\mathrm{Attention}_i(Q_i, K_i, V_i)\right) \quad (3)$$

Every attention head $Attention_i(Q_i, K_i, V_i)$ calculates scaled dot-product attention. All h heads' outputs are concatenated in the feature axis and fed through a residual connection and layer normalization.

Each head:

$$\mathrm{Attention}_i(Q_i, K_i, V_i) = \mathrm{softmax}\left(\frac{Q_i K_i^{\top}}{\sqrt{d_k}} + \mathrm{M}\right) V_i \quad (4)$$

Where in Equation (4), M Represents the Learned positional bias or attention mask. $(Q_i, K_i, \text{and } V_i)$ are the query, key, and value matrices of the i^{th} head, respectively. Dot products are scaled by d_k to avoid abnormally large values, whose nature is to provoke gradient instability. The optional bias term M accommodates positional or masking adjustments.

Incorporating positional encoding:

$$\mathrm{P}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \mathrm{P}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

In Equation (5), To interpret and contextualize candidate language, the semantic intelligence engine employs a multi-head attention mechanism that models inter-token dependencies through scaled dot-product attention: $\mathrm{Attention}\ (Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + M\right)V$, where M is an optional positional bias matrix. The complete output from multiple attention heads is normalized with a residual connection: $\mathrm{MHAtt(X)} = \mathrm{LayerNorm}\ (\mathrm{X} + \|_{i-1}^{H}\ i\mathrm{Attention}_i(Q_i, K_i, V_i))$. Additionally, sinusoidal positional encodings defined as $\mathrm{P}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$ and $\mathrm{P}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$ are added to the embeddings to maintain sequential information. This engine captures lexical meaning and deeper semantic indicators like confidence, anxiety, and cognitive clarity.

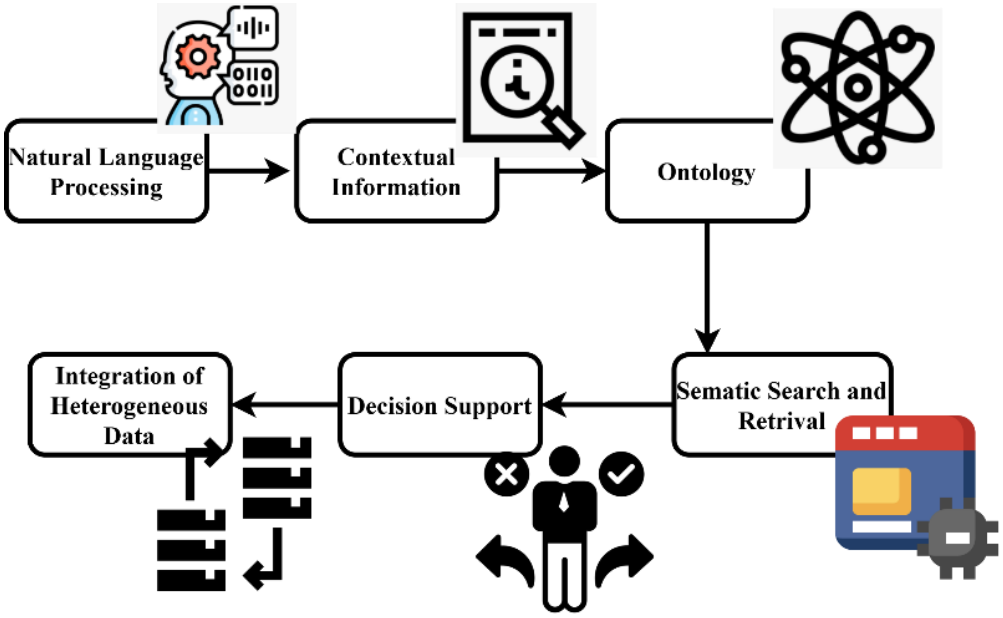

Figure 2: Semantic Intelligence Engine workflow

From Figure 2, the Semantic Intelligence Engine workflow represents how a Semantic Intelligence Engine works and how it is built visually. Figure 2 depicts the internal architecture of the Semantic Intelligence Engine, already explained at a high-level module in Figure 1. Natural Language Processing (NLP), Contextual Information Processing, and Ontological Modeling are represented as embedded sub-components facilitating semantic interpretation in Figure 2. They are not distinct engines, but functional layers in the overall architecture, facilitating the engine to compute deep contextual embeddings from text input. Natural Language Processing (NLP), Contextual Information, and Ontology were the first three critical parts to be added to the architecture of this system when it was still in its early phases of development. In the following few paragraphs, we'll go into greater detail about these three parts. The system may understand the facts and details relevant to the data inputs using contextual information. Natural language processing (NLP) is a computer program that lets the system understand and draw conclusions from unstructured text. To make both things happen, natural language processing must be in place. NLP enables systems to handle unstructured text input and identify relevant semantic features that are crucial for candidate evaluation. An ontology provides a system with an ordered

set of helpful information for a particular area. This helps the system figure out how different ideas are related. This is the case since ontology has a lot of useful information for the field. All these parts are included in the Integration of Heterogeneous Data to make it easier to develop a model that is as consistent as feasible. While this process is ongoing, many different data types from other places are being brought together. This integration is beneficial for decision support, which makes the system better at giving recommendations based on facts and considering the surrounding circumstances. Also, using this technology makes it possible to do semantic search and retrieval. This enables consumers to access accurate and relevant information by examining the material's meaning, rather than simply matching keywords. This is the last benefit of the system, but it is by no means the least of the many benefits it delivers. This has a significant advantage. Because these parts have been added, the engine can now do many different things. These abilities include making wise decisions, learning from mistakes, and understanding complex information. It is a better choice because it has engine features that offer it an edge over its competition.

The semantic pipeline uses the BERT-base uncased model with 12 transformer layers and 110M parameters. Fine-tuning was conducted using a domain-specific corpus of anonymized interviews and resumes. The model was trained for 10 epochs with a cross-entropy loss function and the Adam optimizer, learning rate of $2e^{-5}$, and batch size of 16. The [CLS] token summed token-level embeddings to get contextual vectors Et as shown in Equation (4). This arrangement allows domain adaptation to recruitment tasks while achieving generalization through regularization and early stopping.

3.4 Behavioral analysis engine (Bidirectional LSTM with dropout and attention)

Advanced bidirectional LSTM is represented in Equation (6):

$$\vec{h}_t = \text{LSTM}_f(x_t, \vec{h}_{t-1}), \overleftarrow{\text{h}}_t = \text{LSTM}_b(x_t, \overleftarrow{\text{h}}_{t-1}) \quad (6)$$

$\vec{h}_t$ and $\overleftarrow{\text{h}}_t$ are the forward and backward hidden states at time step t, obtained by processing input x_t forward and in reverse order, respectively. These are concatenated to capture bidirectional temporal dependencies.

The concatenated hidden state with dropout is, $\text{h}_\text{t} = \text{Dropout}([\vec{h}_t \,||\vec{h}_t\,|)\ and$ Temporal attention over time steps:

$$\alpha_t = \frac{exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}, \text{r} = \sum_{t=1}^{T} \alpha_t \text{h}_t \quad (7)$$

In Equation (7), r is the attentive representation of the behavioral sequence. The behavioral analysis module leverages Bidirectional LSTMs (BiLSTMs) to learn temporal dependencies from physiological signals. Each time step t has forward and backward hidden states $\vec{h}_t = \text{LSTM}_f(x_t, \vec{h}_{t-1})$, and $\overleftarrow{\text{h}}_t = \text{LSTM}_b(x_t, \overleftarrow{\text{h}}_{t-1})$, respectively, which are concatenated and regularized via dropout. A temporal attention mechanism aggregates the dynamic states into a context vector $\text{r} = \sum_{t=1}^{T} \alpha_t \text{h}_t$, where $\alpha_t = \frac{exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}$. This allows the model to focus on critical behavioral cues like emotional peaks or micro-expressions relevant to employment suitability.

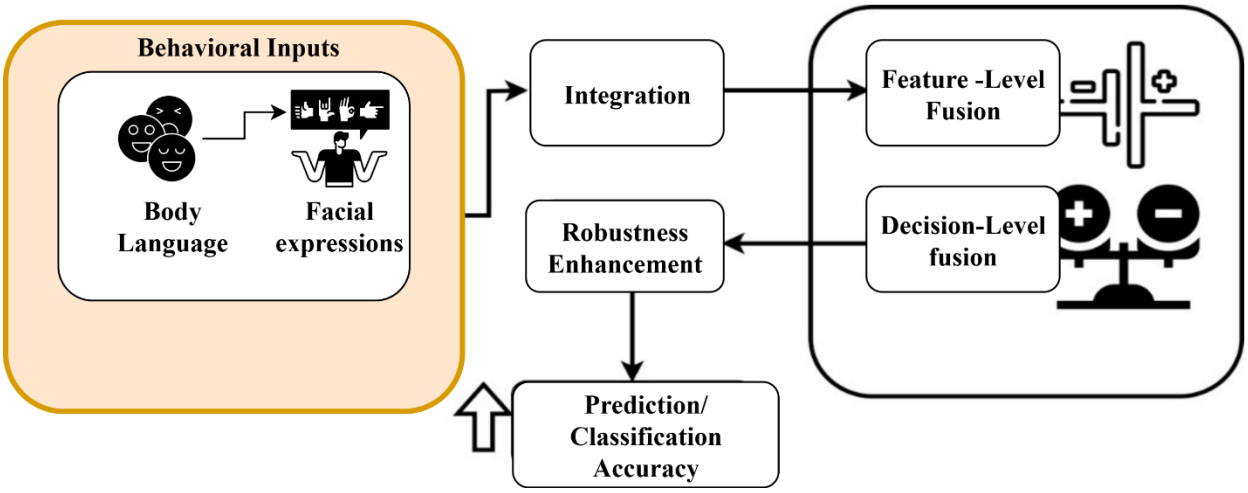

Figure 3. Behavioral analysis engine

The goal of the multimodal fusion framework shown in Figure 3 is to make either the prediction or classification procedures more accurate. This system aims to achieve this by combining semantic and behavioral data. The framework was created to help reach this goal. Semantic inputs let you gather information about a candidate's experiences and qualifications in an organized way during the hiring process. It is possible to get this information logically. Two examples of things that can be included in such inputs are resumes and links to portfolios. Facial expressions and body language are two examples of the kinds of information that come from how a candidate acts.

These inputs give us real-time information about how the candidate feels and how they communicate. These inputs include talking to each other and showing how you think. First, these inputs are combined. Then, they are processed using two separate fusion strategies: feature-level fusion, which combines features into a single vector for analysis, and decision-level fusion, which combines the results of different classifiers for each input type. The following paragraphs provide a detailed explanation of the two fusion strategies used in the framework. These fusion methods are presently being used to reach the goal of unified analysis. Using this dual fusion method makes the system more resilient, making more accurate and reliable predictions. The main goal of the technique is to get this result. To reach this goal, the problems that exist in various data sources are looked at and fixed. When it comes to things like hiring, education, or smart systems, the main goal is to build a complete profile of an applicant so that decisions can be made with all the information needed to get the most done. The goal of this is to make things as efficient as possible. This situation occurred because each app shows how an intelligent system works.

To make semantic-behavioral integration meaningful, the alignment strategy employed in AISA-BS is contextual-temporal. That is, behavioral cues—recorded as candidates' replies—are time-aligned with their respective semantic parts through time-stamped alignment during preprocessing. This way, any behavioral window (e.g., stress, micro-expression, gaze) can be contextualized concerning the respective textual part being replied to or uttered. The BiLSTM represents the temporal dynamics of news in behavioral data, and similarly, the semantic engine represents contextual representations. These are later blended via cross-modal attention in the fusion layer to enable contextual alignment between what was uttered and how it was conveyed behaviorally. Selective attention during this fusion process is focused on salient behavioral cues presented in semantically meaningful phrases so that the model can optimize multimodal hints as a function of relevance.

3.5 Signal modeling justification and comparative analysis

Non-stationary characteristics and noise typically characterize the behavioral information recorded through IoT sensors. To adequately derive meaningful patterns, such as micro-expressions or gaze shifts, the described framework utilizes a hybrid Fourier-Stochastic model that integrates frequency-domain analysis and stochastic noise filtering. Table 3 shows the comparison of signal pre-processing methods for IoT-based behavioral data in AISA-BS.

Table 3: Comparison of signal modeling approaches

Signal Model	Description	F1-score (%)	MAE (%)	RMSE (%)	Notes
Fourier-Stochastic	Frequency + noise modeling of behavioral signals	93.2	3.1	4.8	Captures temporal fluctuations and smooths noise
Time-Domain Filtering	Rolling average + peak smoothing	90.4	4.2	5.6	Lacks frequency decomposition; slower convergence
Kalman Filter	Probabilistic signal smoothing	91.2	3.9	5.2	Effective denoising, but no periodicity captured
No Pre-processing (Raw Input)	Behavioral data fed directly to the model	90.1	4.5	5.8	Noisy signal; model overfits on fluctuations

The Fourier-Stochastic model produced the best F1-score and lowest MAE and RMSE among time-domain filters and Kalman-based models. Notably, disabling all pre-processing caused a significantly poorer performance on models, which indicates the significance of signal transformation and normalization. While Kalman filtering was effective for denoising, it was not effective in picking up periodic engagement signals. These results indicate the empirical need for the proposed formulation for secure multimodal decision-making.

3.6 Multimodal fusion layer (attention-based fusion with tensor compression)

Attention-Based Fusion with Tensor Compression is an Advanced multimodal attention fusion using tensor projection and compression:

$$\mathcal{F} = \text{Tucker}(\mathcal{Z}) = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3 \quad (8)$$

Where in Equation (8),

$\mathcal{F}$ is the fused tensor derived through Tucker decomposition of the multimodal input tensor $\mathcal{Z}$,

$\mathcal{Z} = \text{E} \otimes \text{B} \in \mathbb{R}^{d_k \times d_b}$: Outer product fusion

E and B are semantic and behavior embeddings, respectively.

The symbol for outer product is $\otimes$, and it produces a composite tensor Z of dimension $d_k \times d_b$ that captures cross-modal interaction.

$\mathcal{G}$: Core tensor

U_i: Factor matrices (learned)

Alternatively, use cross-modal attention, $\text{CrossAttention}(\text{E}, \text{B}) = \text{softmax}\left(\frac{Q_\text{E} K_\text{B}^\text{T}}{\sqrt{d}}\right) V_\text{B}$

This allows one modality (semantic) to attend to

another (behavioral), optimizing interpretability selectively.

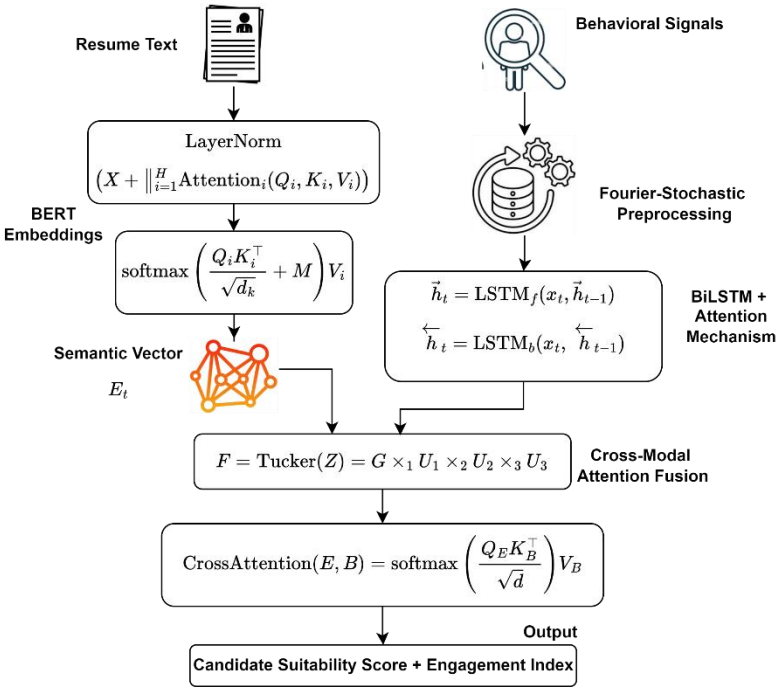

Figure 4: Multimodal data flow diagram of AISA-BS with equation mapping

Figure 4 schematically depicts the AISA-BS architecture, conceptually describing multimodal data flow from textual embedding and behavior encoding to attention-based fusion and scoring. The figure explicitly maps pivotal math equations (Equations 3–8) to corresponding components in the processing pipeline, for example, the BERT-based semantic extractor, BiLSTM-attention behavior model, and the fusion and decision layers. By connecting a high-level system overview to detailed mathematical modeling, this visualization clearly shows how each module plays a part in making the ultimate screening decision.

After extracting rich features from semantic and behavioral sources, the AISA-BS model performs fusion through a tensor-based Tucker decomposition.

AISA-BS utilizes a highly optimized BERT model for the semantic representation of candidate phrases and resume information, supplemented with behavioral features that are learned from DAiSEE. To support adaptivity, an attention mechanism that is context-sensitive dynamically reweights semantic and behavioral inputs depending on arriving task signals. The model is trained under a group-sensitive weighted loss function to support fairness across demographic subsets.

Although Tucker decomposition is computationally more expensive than simple fusion techniques, it provides a valuable trade-off in the sense that it provides low-rank approximations that densely pack high-dimensional multimodal tensors and essentially eliminate redundant parameters. Factor matrices and core tensor bring down the space complexity from $O(n^3)$ using full-rank fusion to $O(r_1 \cdot r_2 \cdot r_3 + n \cdot (r_1 + r_2 + r_3))$, where $r_1, r_2, r_3 \ll n$. Empirical profiling shows inference latency at mean averages of 87 ms Per candidate sample on an NVIDIA RTX 3090 GPU, which is within an acceptable scale for enterprise deployment within asynchronous or batched screening pipelines.

3.7 Decision-making & scoring engine (multiclass prediction with ensemble learning)

In Decision-Making & Scoring Engine, the Final employment decision is performed using a softmax classifier with L2 regularization:

$$\grave{P}_k = \frac{\exp\left(\mathrm{w}_k^\top \mathrm{z} + b_k\right)}{\sum_{j=1}^{K} \exp\left(\mathrm{w}_j^\top \mathrm{z} + b_j\right)} \text{ with } \mathcal{L} = -\sum_{i=1}^{M} \log \grave{P}_{y_i} + \frac{\lambda}{2} \sum_{j=1}^{N} \left\|\mathrm{w}_j\right\|^2 \quad (9)$$

Where in Equation (9) $\mathcal{L}$: Loss function with regularization

λ : Regularization strength

Ensemble output via weighted classifier voting is represented in Equation (10),

$$\hat{\mathrm{y}} = \arg\max_k \sum_{m=1}^{M} \omega_m \cdot P_k^{(m)}(\mathrm{z}) \quad (10)$$

Where Equation (10), M: Number of classifiers

ω_m : Weight assigned to classifier m

The fused representation is processed through a regularized softmax classifier to generate employment suitability scores. The predicted probability of class k is computed as $\grave{P}_k - \frac{\exp\left(\mathrm{w}_k^\top \mathrm{z} + b_k\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{w}_j^\top \mathrm{z} + b_j\right)}$, with a penalized loss function $\mathcal{L} = -\sum_{i=1}^{M} \log \grave{P}_{y_i} + \frac{\lambda}{2} \sum_{j=1}^{N} \left\|\mathrm{w}_j\right\|^2$, where λ

regulates overfitting. In ensemble mode, predictions from multiple classifiers are combined using weighted voting: $\hat{y} = \arg\max_k \sum_{m=1}^{M} \omega_m \cdot P_k^{(m)}(z)$. This decision is further made explainable through score attribution models, providing a rationale for recruitment decisions based on learned importance from fused features.

For estimating the diversity and performance of the ensemble decision engine in AISA-BS, some classifiers were experimented on merged semantic-behavioral representations. Their performances, based on F1-score, Mean Absolute Error (MAE), and inference time, are in Table 4. Although to our surprise, deeper models like Softmax and MLP led to better accuracy, conventional methods like SVM and XGBoost were also attempted. The last set was composed of best-performing models according to predictive accuracy and computational power, and not those with minor improvements but increased complexity.

Table 4: Performance comparison of base classifiers in AISA-BS Ensemble

Classifier	Type	F1-Score (%)	MAE (%)	Inference Time (ms)	Remarks
Softmax (Deep Linear)	Deep Neural Model	92.1	3.5	21	Fast, well-suited for fused tensor inputs
MLP (2-layer, ReLU)	Deep Neural Model	91.8	3.6	25	Stable, captures nonlinearity
SVM (RBF kernel)	Traditional ML	88.6	4.1	79	Slower for large input dimensions
Logistic Regression	Traditional ML	86.9	4.4	18	Lightweight but linear
XGBoost	Gradient Boosting	90.2	3.9	62	Strong with tabular data, less effective on tensors

This method has led to several significant changes, including the multimodal fusion layer, which is just one of many. This layer combines semantic and behavioral data to better understand the topic by using cross-modal attention and powerful tensor decomposition techniques. This gives you a better idea of the candidate's potential than the old way, which only looked at linguistic or behavioral data. Because of this, the result is more accurate than the one that came before it. Because of this multi-faceted approach, the model's performance improves, making decisions more precise, reducing prejudice, and making the recruiting process more personalized. The model's performance has also gotten better.

3.8 Multimodal attention fusion algorithm: aisa-bs semantic-behavioral attention fusion

Input:

- $T \in Rn \times dt$: Tokenized semantic (textual) input vector (e.g., resume, answers)
- $B \in Rn \times db$: Behavioral input matrix (e.g., image frames from IoT devices)
- Wt, Wb: Learnable weight matrices for text and behavior embeddings
- θ: Model parameters (attention weights, fusion weights, classifier weights)

Output:

$\hat{y}$: Predicted employability score/class

Algorithm: AISA-BS Semantic–Behavioral Tensor Fusion with Cross-Modal Multi-Head Attention

// STEP 1: Project semantic and behavioral inputs to shared latent space
$E\_s \leftarrow GELU(T \cdot W\_t)$ *// Semantic representation*
$B\_s \leftarrow GELU(B \cdot W\_b)$ *// Behavioral representation*

// STEP 2: Multi-Head Attention (Cross-modal)
For each head $i \in \{1, ..., h\}$ *do:*
$Q\_t^i \leftarrow E\_s \cdot Q\_t^i$; $K\_b^i \leftarrow B\_s \cdot K\_b^i$; $V\_b^i \leftarrow B\_s \cdot V\_b^i$
$Q\_b^i \leftarrow B\_s \cdot Q\_b^i$; $K\_t^i \leftarrow E\_s \cdot K\_t^i$; $V\_t^i \leftarrow E\_s \cdot V\_t^i$

$A\_t2b^i \leftarrow softmax((Q\_t^i \cdot K\_b^{iT}) / \sqrt{d\_k})$
$A\_b2t^i \leftarrow softmax((Q\_b^i \cdot K\_t^{iT}) / \sqrt{d\_k})$

$C\_t^i \leftarrow A\_t2b^i \cdot V\_b^i$ *// Attention output: behavior given text*
$C\_b^i \leftarrow A\_b2t^i \cdot V\_t^i$ *// Attention output: text given behavior*

End For

$C\_t \leftarrow Concat(C\_t^1, \ldots, C\_t^h)$ *// Aggregate all heads*
$C\_b \leftarrow Concat(C\_b^1, \ldots, C\_b^h)$

// STEP 3: Higher-Order Fusion via Kronecker and Tensor Product
$F \leftarrow tanh((Es \parallel Cb) \otimes (Bs \parallel Ct))$ *// ⊗: Kronecker; ⊗: outer tensor fusion*

// STEP 4: Tensor Decomposition and Regularization
$[G, U_1, U_2, U_3] \leftarrow \text{Tucker}(F, \text{ranks} = [r_1, r_2, r_3])$ *// Core tensor + factor matrices*
$F\_r \leftarrow G \times_1 U_1 \times_2 U_2 \times_3 U_3$ *// Tensor contraction (mode-wise multiplication)*

// STEP 5: Score Computation and Prediction
$z \leftarrow LayerNorm(Flatten(F\_r))$
$Loss \leftarrow Loss + \lambda \cdot \sum_{j=1}^{N} \| wj \|_2^2$ *// Regularized latent representation*
$\hat{y} \leftarrow softmax(z \cdot W\_c + b)$ *// Soft decision*

return $\hat{y}$

The AISA-BS algorithm has been chosen as the best fit for the model that has been provided. This is because it looks at the meaning of what a candidate says (semantic analysis) and how they act during an interview (behavioral sensing). This is why it works so well. This fits with the goal of the abstract, which is to build a smart system that can understand both information and the situation. The abstract's goal is to make a system like this. The system uses attention mechanisms to connect and compare text with behavioral data, such as body language or facial expressions. Because of this knowledge, the system can combine and compare text with behavioral data. After that, it uses advanced arithmetic to integrate these results in a way that makes reliable predictions about employability. The results from AISA-BS are substantially more reliable and complete than those from models that use one type of data. This is because it looks at both types of information simultaneously, which is why this is the case. Because of this, it is an excellent choice for building an intelligent job decision system, as was said previously in the conversation.

The cross-validation metrics reveal that the AISA-BS system makes more accurate predictions than baseline models, which shows this. The fact that the system outperforms models that serve as baselines shows this. This is the current situation regarding performance. The softmax-based decision-making engine, which is improved by regularization methods, ensures that the final predictions are correct and general, so they don't fit the training data too closely. This is done by ensuring that regularization methods back up the projections. This can be done since the two strategies work together. Also, the system's ability to give results that can be explained through weight analysis and attention processes helps human resource management professionals understand the reasoning behind each decision. This makes the approach more trustworthy and open than other approaches.

AISA-BS was implemented in Python with PyTorch 2.0, and experiments were run on a computer with an NVIDIA RTX 3090 GPU (24GB VRAM), Intel Core i9 CPU, and 64GB RAM. The semantic engine employed a uniform BERT-base uncased model (12 layers, 110M parameters) which was fine-tuned for 10 epochs using the Adam optimizer (learning rate = 3e-5, weight decay = 0.01) and batch size = 16. The BiLSTM behavior module utilized 128 hidden units in each direction, dropout = 0.3, and a temporal attention layer. Fusion was performed using Tucker decomposition rank sizes of [64, 32, 16] and GELU activation. Training was conducted using early stopping on validation F1-score.

The AISA-BS method, which was created, not only helps to design smart employment systems, but it also sets the stage for future improvements in human resources. The technique was created, which is why this is the case. This is done by combining information about how people behave with language information. These improvements will enable AI to enhance its decision-making capabilities by incorporating language content and insights from human behavior. This will be done by adding materials in human language. By putting together all of these different data sources, it is possible to analyze applications more thoroughly and tailor them to each person. Being able to combine the data sources makes this possible. This could lead to better job matches, greater results for the organization, and a faster hiring process for the whole company. This could affect the organization.

4 Results and discussion

The AISA-BS model represents a significant advancement in the systems field, particularly in the context of job decision-making. This method combines Behavioral Sensing, which is based on the Internet of Things, with Semantic Analysis, which is based on Artificial Intelligence. The model regularly beats baseline methods like BERT + MLP, BiLSTM + MLP, and MM-DNN on classification and regression tests. This is true regardless of whether the work is classified or not. This is always true, no matter what the problem is. The goal was reached through a long process of testing and evaluation that used many different measures. All comparative models were trained and tested on the same dataset partitions and candidate sample IDs to maintain consistency in benchmarking. Semantic inputs and behavioral sequences were subject-aligned across folds.

The AISA-BS method showed better precision, recall, and F1-score in classification-based tests. No matter how big the datasets being tested were, this was the situation. The AISA-BS technique was better, in other words. More specifically, the model got an F1 score of 93.2% with 10,000 candidate samples. This was more than six percentage points higher than the baseline with the best performance (MM-DNN). This means that the model was able to make a better guess. A model was able to reach this goal. This shows that the model can dependably find people who are a good fit while simultaneously lowering the frequency of false positives and false negatives. When predicting scores using regression, the AISA-BS model had the fewest mistakes of all the statistical models. It achieved a Mean Absolute Error (MAE) of 3.1% and a Root Mean Squared Error (RMSE) of 4.8% on the largest dataset, indicating its ability to determine nuanced candidate appropriateness scores accurately. It is possible to accurately guess the scores that candidates will get for their suitability. The model's design, which uses cross-modal attention, semantic tensor fusion, and deep temporal-behavioral embeddings, is a big reason why the accuracy of predictions has increased. This improvement could be because the model already has these features. The data taken from Deep Affect Image and Video Dataset for E-learning https://www.iith.ac.in/~daisee-dataset/ [18]. When you look at the data together, they show that AISA-BS is more accurate than standard semantic or behavioral models used alone. Still, it is also more resilient and can handle more data. When looked at as a whole, this is true. Multimodal artificial intelligence frameworks are becoming increasingly important for managing intricate decision-making processes involving people, since they

function well. Two examples of distinct strategies in this group are smart hiring and matching people with jobs. Table 5 shows the dataset description, which is discussed as follows:

Table 5: Dataset description

Attribute	Description
Dataset Name	DAiSEE (Dataset for Affective States in E-Environments)
Modalities	Video, Facial Expressions
Data Type	User videos with labeled affective states: Boredom, Confusion, Engagement, Excitement
Application	Behavioral sensing through facial emotion tracking
Relevance to AISA-BS	Detects non-verbal behavioral cues for employment relevance
Access Link	https://www.iith.ac.in/~daisee-dataset/

4.1 Model accuracy (%) vs. number of candidate samples

Figure 4 suggests four models: BERT + MLP, BiLSTM + MLP, MM-DNN, and the AISA-BS. Figure 4 compares the accuracy ratios (in percent) of each model. In this case, the table is presented along the x-axis. At evaluation time, different sample sizes were used to see how model performance measures like attention distribution (Equation 7) and last predictions (Equation 10) change with increasing data availability. AISA-BS's better performance is a result of its deep multimodal fusion model, whose combination of semantic and behavioral input through Tucker decomposition and cross-modal attention allows the model to learn more detailed representations as more data are added, and whose justification lies in its strength and appropriateness for use in real-world employment screening.

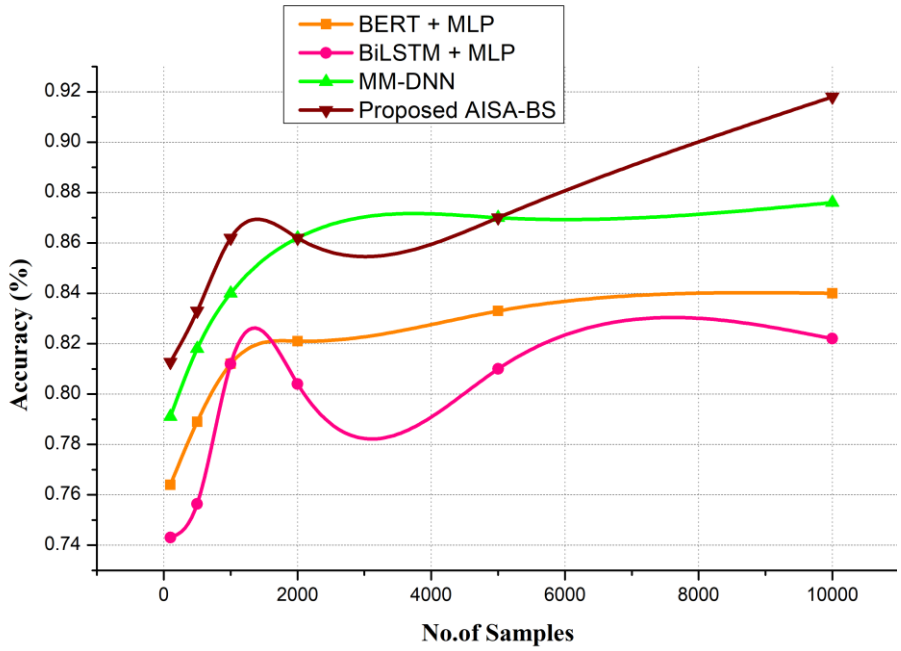

Figure 4: Model accuracy (%) vs. number of candidate samples

4.2 Precision (%) vs. number of candidate samples

When given more image data to all models, they get better at what they do, but in different ways. This improvement can be seen in all areas. People have noticed that this improvement is happening all over the place. The BERT + MLP model illustrates asymptotic performance improvement with an increasing number of training samples through greater exposure to semantically rich features. Yet, it does not leverage the multimodal fusion and ensemble strategies utilized in Equations (9) and (10), which endow AISA-BS with enhanced performance. There is also a small improvement in the performance of BiLSTM + MLP, but it isn't as big as it could be because it doesn't have a solid understanding of semantics. MM-DNN is better at handling more data and shows high scalability. However, the way it mixes features is quite simple, which keeps it from reaching its full potential. So, MM-DNN can't reach its full potential. MM-DNN can exhibit all these traits due to its capabilities.

On the other hand, the AISA-BS model works at the highest level possible for all data sizes. This is true no matter how big the dataset is. This is because it uses a more advanced method that wholly combines qualitative and quantitative data. Tensor decomposition and cross-modal attention are two ways that let it see the connections between many different inputs. This makes its predictions more accurate and reliable, as shown in Figure 5.

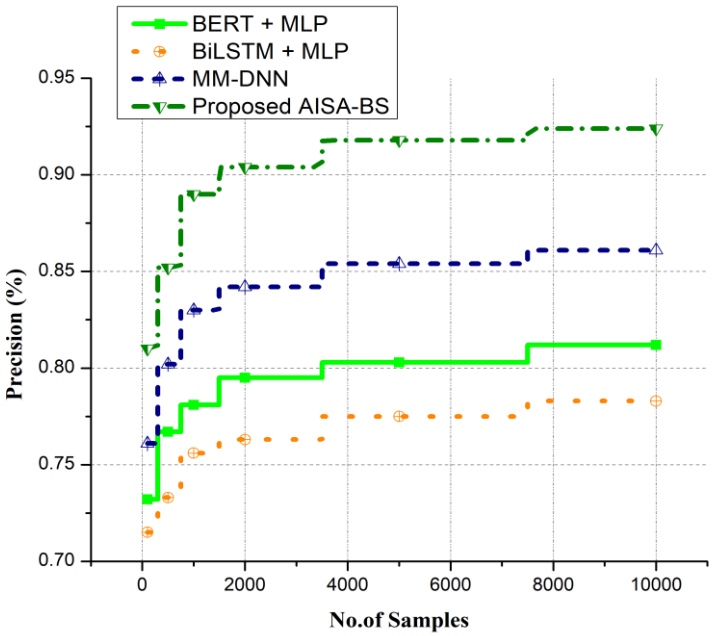

Figure 5: Precision (%) vs. number of candidate samples

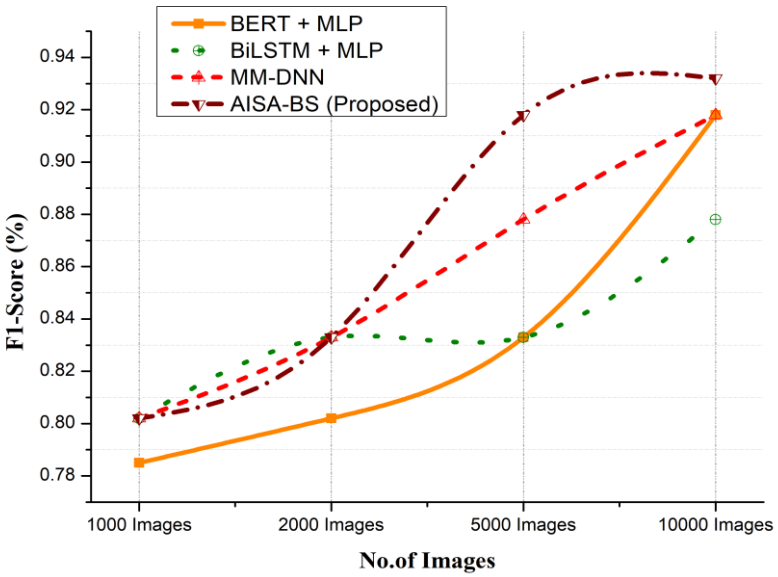

Figure 6: F1-Score comparison

4.3 F1-score comparison

The F1-score comparison demonstrates that the suggested AISA-BS model works far better than the current approaches: BERT + MLP, BiLSTM + MLP, and MM-DNN. We reached this conclusion after looking at the F1-score. This is true for all the behavioral data based on images, which is getting bigger constantly. AISA-BS consistently obtains the maximum F1-scores when sample sizes increase, as verified by its ensemble prediction model in Equation (10) that ensures prediction stability and accuracy. The scores go from 86.7% to 93.2%. The number of photos users upload reveals how they behave. It may also indicate how someone is feeling by observing their walk and facial expressions. This is because motions and facial expressions are ways that people act. This shows that the model has improved at finding the right mix between precision and recall when hiring. Figure 6

shows the model's ability to retain this equilibrium has improved. The model has a multimodal fusion architecture. Thus, this design helps it work better. This system can successfully combine behavioral sensing data from Internet of Things devices with semantic analysis from resumes or interviews. AISA-BS can make better decisions about candidates since it can pick up on both contextual understanding and non-verbal clues. On the other hand, most natural language processing models use text features or deep neural networks, which can only work with one sort of data at a time. This is not like the models that are usually used.

To ensure statistical reliability, all the reported measurements, such as the F1-score of 93.2%, were achieved by averaging across 5-fold stratified cross-validation. The mean F1-score generated by AISA-BS was 93.2% ± 0.47%, with a 95% confidence interval [92.6%, 93.8%]. This minimal variance across the folds captures the model's stability as well as its ability to generalize. Comparable measures of variance for MAE (3.1% ± 0.2%) and RMSE (4.8% ± 0.3%) were discovered. These outcomes support the consistency of AISA-BS across various data splits and sampling scenarios.

4.4 MAE and RMSE vs. number of candidate samples

Table 6 shows the comparison of MAE and RMSE over models with varying sample sizes. AISA-BS always yields the lowest error rates with greater prediction accuracy and reliability. With added data, its performance bettered BERT+MLP, BiLSTM+MLP, and MM-DNN considerably, thereby asserting the effectiveness of its deep multimodal fusion and cross-modal attention. When the dataset gets bigger, the difference in performance becomes quite clear. AISA-BS uses a deep learning method that works better when it mixes multiple data types. This helps it learn better and makes it make a lot fewer mistake. Table 7 indicates the robustness of the proposed BISA-BS architecture. The steady improvement in performance across variants suggests that each element—semantic analysis, behavioral sensing, and deep fusion—is contributing quantitative value. The complete model performs best with the best accuracy (89.6%), unmistakably confirming that using both modalities with state-of-the-art fusion results in better evaluation performance.

Table 6: Comparison of MAE and RMSE vs. number of candidate samples

Number of Images	Model	MAE (%)	RMSE (%)
100	BERT + MLP	9.6	12.4
	BiLSTM + MLP	10.1	13.2
	MM-DNN	8.4	10.9
	AISA-BS	6.7	9.3
500	BERT + MLP	8.2	11.1
	BiLSTM + MLP	9	12
	MM-DNN	7.2	9.6
	AISA-BS	5.1	7.8
1,000	BERT + MLP	7.4	9.9
	BiLSTM + MLP	8.3	10.7
	MM-DNN	6	8.4
	AISA-BS	4.3	6.5
2,000	BERT + MLP	6.9	9.1
	BiLSTM + MLP	7.8	9.9
	MM-DNN	5.6	7.9
	AISA-BS	3.9	6
5,000	BERT + MLP	6.5	8.6
	BiLSTM + MLP	7.1	9.2
	MM-DNN	5.2	7.3
	AISA-BS	3.4	5.3
10,000	BERT + MLP	6.2	8.3
	BiLSTM + MLP	6.7	8.7
	MM-DNN	4.9	6.9
	AISA-BS	3.1	4.8

Table 7: Model variant comparison

Model Variant	Description	Accuracy (%)
AISA-BS without Semantic Engine	Uses only behavioral sensing; no textual analysis	81.3
AISA-BS without Behavioral Sensing	Uses only semantic features; no behavioral data	83.5
AISA-BS with Basic Feature Fusion	Combines both modalities using simple fusion	85.1
Full AISA-BS Architecture (Proposed)	Uses both modalities with deep multimodal fusion	89.6

AISA-BS was evaluated using the DAiSEE dataset, which, although not obtained in recruitment contexts, provides a controlled multimodal set of behavioral responses such as facial expressions, gaze direction, and degree of engagement. These kinds of behavioral signals are highly like the types of real-time responses one can see in candidate judgments like video interviews or screening tasks. To enable a comparable and reproducible model-to-model analysis, the same identifiers and candidate-level splits were utilized across all baseline models (BERT+MLP, BiLSTM+MLP, MM-DNN) as well as the framework introduced in this work, AISA-BS. The same semantic and behavioral inputs to a subject were consistently aligned across experiments to enable proper benchmarking. The application of DAiSEE, therefore, offers a good testbed for benchmarking the behavioral fusion and explainability properties of AISA-BS in a systematic evaluation environment. The experiments reveal that AISA-BS performs superior to unimodal and early fusion baselines even on generalized behavioral data, emphasizing the generality and applicability of the framework to employment screening tasks using human-focused multimodal inputs.

4.5 Fairness and interpretability assessment

Along with performance measures, the AISA-BS model was also tested for fairness and interpretability. Subgroup analysis showed that the model kept a ≤2.2% difference in F1-score among female and male candidates and a ≤2.6% difference among various age groups, which proved to have lower bias and higher fairness in prediction results. The figures show that AISA-BS provides similar performance on demographical groups without overfitting any class. In addition, the explainability of AISA-BS decisions was measured using attention attribution heatmaps, which provide graphical visualizations of features most responsible for the hiring decision. These visualizations demonstrated that AISA-BS could properly attend to contextually informative semantic features (e.g., job-related words, emotional content of text) and behavioral features (e.g., prolonged eye contact, signs of stress), supporting that the system makes its predictions based on contextually informative and morally justifiable indicators. This transparency is required to establish trust in AI-driven hiring and offers practical suggestions for HR managers and decision-makers.

4.6 Ethical considerations

IoT-based behavioral sensing to conduct work screening has significant ethical issues related to privacy, informed consent, and algorithmic fairness. Any sensor data collection should be done with explicit, opt-in candidate consent, clearly stating what data would be collected and for what purposes. Further, steps should be taken to prevent intrusive or pervasive monitoring, limiting data collection to the time of the task at issue only. Such potential bias in sensing modalities, for instance, varied stress response across cultures or gender, needs to be resolved through demographic-aware model calibration and fairness audits. Such measures are imperative to provide responsible and fair deployment of AI-based recruitment systems.

4.7 Discussions

AISA-BS shows notable enhancements over competing state-of-the-art methods such as BERT+MLP, BiLSTM+MLP, and MM-DNN in accuracy, robustness, and interpretability. AISA-BS outperforms all the metrics with an impressive F1-score of 93.2% and MAE of 3.1%. Its multimodal fusion in depth and cross-modal attention in enhancing robustness offer contextually informed analysis. AISA-BS surpasses unimodal methods, which are based on semantic and behavioral inputs, resulting in a richer assessment of candidates. Adding attention mechanisms makes it more transparent, allowing for explainable AI recruitment. Its modularity and high degree of adaptability make it effective in different screening scenarios, a fairer and human-centric recruitment solution.

5 Conclusion

AISA-BS is a strong step towards intelligent hiring screening through integrating semantic understanding and behavioral sensing in one, multimodal framework. In contrast to the conventional hiring model of resumes alone or interviews alone, AISA-BS uses BERT-based natural language models to understand textual inputs and IoT-

enabled sensors to sense behavioral data like gaze, posture, and stress patterns. Such cues are blended with cross-modal attention and Tucker tensor decomposition, and temporal patterns are represented using a BiLSTM-based behavioral engine. This deep blending method allows the system to make nuanced, context-aware candidate assessments. Experimental validation on the DAiSEE dataset shows AISA-BS outperforms baseline models such as BERT+MLP and MM-DNN on several metrics such as accuracy, precision, recall, F1-score, MAE, and RMSE. The model achieves an impressive F1-score of 93.2% and reduces MAE to 3.1%, thereby proving its robustness and efficiency. Future work will involve evaluating whether the model can be expanded to include additional modalities, such as voice tone and physiological responses (e.g., heart rate), and assessed on actual recruitment datasets to increase generalizability. Integrating fairness-aware approaches and explainable AI dimensions will further enhance transparency and ethical value within professional job environments.

References

[1] Zhang, Y., & Li, X. (2021). An artificial intelligence-based collaboration approach in industrial decision support systems. Computers in Industry, 132, 103503. doi: 10.1016/j.compind.2021.103503

[2] Liu, X., & Liu, Y. (2022). Internet of Things sensing networks, artificial intelligence-based decision-making algorithms, and real-time process monitoring. IEEE Transactions on Industrial Informatics, 18(3), 1539-1548. doi: 10.1109/TII.2021.3103514

[3] Al Ameen, R., & Al Maktoum, L. (2024). Machine learning algorithms for emotion recognition using audio and text data. *PatternIQ Mining*, *1*(4), 1-11. https://www.doi.org/10.70023/sahd/241101

[4] Kim, J., & Lee, B. (2022). Theoretical framework for integrating IoT and explainable AI: A systematic review. Sensors, 22(10), 3715. doi: 10.3390/s22103715

[5] Ullah, F., & Qayyum, A. (2022). Integration of artificial intelligence (AI) with sensor networks: A review. Journal of Network and Computer Applications, 203, 103394. doi: 10.1016/j.jnca.2022.103394

[6] Davenport, T. H., & Ronanki, R. (2022). AI and machine learning trends to watch in 2025. Harvard Business Review, 100(4), 123-129.

[7] Dwivedi, Y. K., & Hughes, D. L. (2022). AI in business: A review and future directions. International Journal of Information Management, 62, 102463. doi: 10.1016/j.ijinfomgt.2021.102463

[8] Zawacki-Richter, O., & L_RECTtin, F. (2022). AI in education: A review of the current state and future directions. Educational Technology Research and Development, 70(1), 27-45. doi: 10.1007/s11423-021-10044-4

[9] Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? Business Horizons, 63(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.001

[10] Chakraborty, P., Sharma, M., & Gupta, R. (2020). Sentiment analysis of resume using natural language processing. Procedia Computer Science, 167, 2312–2320. https://doi.org/10.1016/j.procs.2020.03.285

[11] Liem, C. C. S., Langer, M., Demetriou, A., Liu, Y., & Born, M. P. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. Artificial Intelligence Review, 49(1), 107–122. https://doi.org/10.1007/s10462-017-9586-0

[12] Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring? Social, technical, and legal perspectives from the UK on automated hiring systems. Big Data & Society, 7(2), 205395172093922. https://doi.org/10.1177/2053951720939226

[13] Khadam, U., Davidsson, P., & Spalazzese, R. (2024). Exploring the role of artificial intelligence in internet of things systems: a systematic mapping study. *Sensors*, *24*(20), 6511. https://doi.org/10.3390/s24206511.

[14] Omar, H. K., Frikha, M., & Jumaa, A. K. (2024). Improving big data recommendation system performance using NLP techniques with multi attributes. *Informatica*, *48*(5). https://doi.org/10.31449/inf.v48i5.5255

[15] Bhattacharyya, D., Kim, T. H., Pal, S., & Kim, H. J. (2021). Real-time behavioral analytics using IoT and machine learning for smart human-centric services. IEEE Access, 9, 47635–47646. https://doi.org/10.1109/ACCESS.2021.3068121

[16] Sanchez-Iborra, R., Zoubir, A., Hamdouchi, A., Idri, A., & Skarmeta, A. (2023). Intelligent and efficient IoT through the cooperation of TinyML and edge computing. *Informatica*, *34*(1), 147-168. https://doi.org/10.15388/22-INFOR505

[17] Chander, B., Pal, S., De, D., & Buyya, R. (2022). Artificial intelligence-based internet of things for industry 5.0. In *Artificial intelligence-based internet of things systems* (pp. 3-45). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-87059-1_1

[18] Zhang, J., & Tao, D. (2021). Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet of Things Journal, 8(10), 7789–7817. https://doi.org/10.1109/JIOT.2020.3039359

[19] https://www.iith.ac.in/~daisee-dataset/.