IRSTI 28.23.35

10.51889/2959-5894.2026.93.1.020

T. Sembayev[1], Zh. Karsenbay[1*], D. Alimkhanova[1], A. Sydykov[1]

[1] Astana IT University, Astana, Kazakhstan
[*]*e-mail: 255324@astanait.edu.kz*

PREDICTIVE MODELING OF EMPLOYEE BURNOUT VIA SPEECH ANALYSIS: A SYSTEMATIC LITERATURE REVIEW

Abstract

This paper presents a systematic literature review exploring the intersection of employee burnout and speech analysis, proposing a conceptual multidimensional framework for future predictive modeling. The review adheres to the PRISMA 2020 guidelines and includes searching scientific databases, such as Scopus and PubMed, selecting and summarizing studies linking burnout, including the Maslach Burnout Inventory, and various digital phenotyping elements, such as acoustic-prosodic parameters, speech emotion recognition, and natural language processing results. Addressing the identified methodological gaps, this study outlines a theoretical framework that integrates self-supervised speech representations – specifically models like wav2vec, HuBERT, and WavLM – with emotional features, text indicators, and Organizational Network Analysis to inform future management systems. Model portability, data quality, and practical applicability are discussed separately, including cultural and linguistic specifics and personal data protection requirements in Kazakhstan (e.g., the personal data law and privacy governance approaches). The synthesized findings highlight the potential and limitations of current speech-based AI, providing a roadmap for developing ethically sound, context-aware systems for early burnout detection and preventative interventions.

Keywords: burnout detection, digital phenotyping, organizational network analysis, self-supervised learning, speech analysis.

Т. Сембаев[1], Ж. Кәрсенбай[1], Д. Алимханова[1], А. Сыдықов[1]
[1] Astana IT University, г. Астана, Казахстан

ПРЕДИКТИВНОЕ МОДЕЛИРОВАНИЕ ПРОФЕССИОНАЛЬНОГО ВЫГОРАНИЯ СОТРУДНИКОВ НА ОСНОВЕ АНАЛИЗА РЕЧИ: СИСТЕМАТИЧЕСКИЙ ОБЗОР ЛИТЕРАТУРЫ

Аннотация

В данной статье представлен систематический обзор литературы, исследующий пересечение проблематики профессионального выгорания сотрудников и анализа речи, а также предлагается концептуальная многомерная основа для будущего прогностического моделирования. Обзор выполнен в соответствии с руководящими принципами PRISMA 2020 и включает поиск в научных базах данных, таких как Scopus и PubMed, отбор и обобщение исследований, связывающих выгорание (включая опросник выгорания Маслач) и различные элементы цифрового фенотипирования, такие как акустико-просодические параметры, распознавание эмоций в речи и результаты обработки естественного языка. Для устранения выявленных методологических пробелов в данном исследовании описывается теоретическая база, которая интегрирует самоконтролируемые (self-supervised) представления речи – в частности, такие модели, как wav2vec, HuBERT и WavLM – с эмоциональными характеристиками, текстовыми индикаторами и анализом организационных сетей (Organizational Network Analysis) для информирования будущих систем управления. Отдельно обсуждаются переносимость моделей, качество данных и практическая применимость, включая культурные и лингвистические особенности, а также требования к защите персональных данных в Казахстане (например, закон о персональных данных и подходы к управлению конфиденциальностью). Синтезированные результаты подчеркивают потенциал и ограничения современного ИИ на основе анализа речи, предлагая дорожную карту для разработки этически обоснованных, контекстно-ориентированных систем для раннего выявления выгорания и профилактических вмешательств.

Ключевые слова: выявление выгорания, цифровое фенотипирование, анализ организационных сетей, обучение с самоконтролем, анализ речи.

Т. Сембаев[1], Ж. Кәрсенбай[1], Д. Алимханова[1], А. Сыдықов[1]
[1]Astana IT University, Астана қ., Қазақстан

СӨЙЛЕУДІ ТАЛДАУ АРҚЫЛЫ ҚЫЗМЕТКЕРЛЕРДІҢ КӘСІБИ КҮЙІП КЕТУІН БОЛЖАМДЫҚ МОДЕЛЬДЕУ: ӘДЕБИЕТТЕРГЕ ЖҮЙЕЛІ ШОЛУ

Аңдатпа

Бұл мақалада қызметкерлердің кәсіби күйіп кетуі мен сөйлеуді талдаудың қиылысын зерттейтін әдебиеттерге жүйелі шолу ұсынылған, сондай-ақ болашақ болжамдық модельдеу үшін тұжырымдамалық көпөлшемді негіз ұсынылады. Шолу PRISMA 2020 нұсқаулықтарына сәйкес орындалды және Scopus пен PubMed сияқты ғылыми дерекқорларды іздеуді, күйіп кетуді (соның ішінде Маслачтың күйіп кету сауалнамасын) және акустикалық-просодиялық параметрлер, сөйлеудегі эмоцияларды тану және табиғи тілді өңдеу нәтижелері сияқты цифрлық фенотиптеудің түрлі элементтерін байланыстыратын зерттеулерді іріктеуді және жинақтауды қамтиды. Анықталған әдіснамалық олқылықтарды шешу мақсатында, бұл зерттеу болашақ басқару жүйелерін ақпараттандыру үшін өздігінен бақыланатын (self-supervised) сөйлеу көріністерін – атап айтқанда wav2vec, HuBERT және WavLM сияқты модельдерді – эмоционалдық сипаттамалармен, мәтіндік индикаторлармен және Ұйымдық желілерді талдаумен (Organizational Network Analysis) біріктіретін теориялық негізді сипаттайды. Модельдердің тасымалдануы, деректер сапасы және тәжірибелік қолданылуы, соның ішінде мәдени және лингвистикалық ерекшеліктер, сондай-ақ Қазақстандағы дербес деректерді қорғау талаптары (мысалы, дербес деректер туралы заң және құпиялылықты басқару тәсілдері) жеке талқыланады. Синтезделген нәтижелер сөйлеуді талдауға негізделген қазіргі жасанды интеллекттің әлеуеті мен шектеулерін көрсетеді, ерте күйіп кетуді анықтауға және алдын алу шараларына арналған этикалық негізделген, мәнмәтінге бағытталған жүйелерді дамытуға арналған жол картасын ұсынады.

Түйін сөздер: күйіп кетуді анықтау, цифрлық фенотиптеу, ұйымдық желілерді талдау, өздігінен бақыланатын оқыту, сөйлеуді талдау.

Introduction

Main provisions. This systematic review confirms that specific acoustic markers, such as F0 flattening, and linguistic features, including the absolutist index and pronoun usage, serve as viable passive indicators of occupational burnout. A conceptual multidimensional framework is proposed, theoretically integrating self-supervised speech representations—specifically wav2vec 2.0, HuBERT, and WavLM—with text analysis and Organizational Network Analysis. Key obstacles for implementation include the neutrality paradox in post-Soviet professional cultures and stringent legal requirements, pointing to a need for localized model calibration, edge processing, and strict data minimization in Kazakhstan.

Professional burnout remains one of the most significant threats to organizational effectiveness, service quality, and employee health, particularly in highly stressful industries such as medicine, aviation, call centers, and IT. Traditionally, burnout diagnostics have relied on questionnaires such as the Maslach Burnout Inventory (MBI), which measure emotional exhaustion, cynicism, and a decreased sense of personal accomplishment [1]. However, these instruments rely on self-reporting, conduct measurements only intermittently, and often detect the syndrome in its chronic stage [2].

In recent years, burnout has been perceived not only as an individual psychological phenomenon but also as a significant organizational risk, impacting work quality, staff retention, and process safety. However, a methodological gap remains between theoretical models of burnout and the practice of building predictive systems based on digital data, particularly speech signals [2]. The existing literature shows high variability in the definition of target constructs (burnout, stress, depressive symptoms) and in the methods of their measurement (e.g., through self-report scales), which makes it difficult to compare results and reproduce conclusions.

A significant research gap stems from the lack of a consolidated understanding of which acoustic and linguistic speech markers demonstrate robust associations with burnout measures and under what conditions these associations persist. Much of the research utilizes limited samples, inconsistent

recording protocols, and heterogeneous communication scenarios, leaving the transferability of models to natural work contexts (noise, device diversity, multichannel communication, code-switching) poorly established.

In this sense, an architectural aspect requires special attention. Although modern machine learning approaches, including self-supervised speech representations and ASR-NLP, demonstrate the potential to reduce reliance on labeling and improve feature quality, research lacks a unified understanding of the principles of multimodal integration (acoustics, text, and emotional components) and the requirements for interpretability of results for managerial use. Finally, context-specific organizational settings and implementation remain under research. Particularly, the influence of the structure of interactions in teams (including network characteristics of communications) and the framework of ethical and legal regulation of the processing of personal data are rarely considered as integral elements of the model and validation of such systems, especially when localized for specific jurisdictions.

Therefore, digital phenotyping is an evolving field, shifting from active surveys to the collection of passive data on people's daily behavior and digital traces. In this regard, speech is considered one of the most promising biomarkers of stress and burnout, as its parameters depend on the interaction of cognitive, emotional, and physiological processes and, to some extent, even extend beyond conscious control [3].

The purpose of this literature review is to systematize research on speech and linguistic markers of burnout, as well as the architecture of emotion recognition models based on speech, along with the organizational, cultural, and legal contexts for their application in the workplace. Particular attention is given to synthesizing literature that informs the conceptualization of an ethically sound, speech-based predictive framework localized for the Kazakhstani context. The review is organized thematically. First, the theoretical foundations of burnout and digital phenotyping are considered. Second, speech and linguistic markers of burnout, computational architectures and datasets, organizational and cultural-legal contexts are reviewed. Finally, the literature discusses the potential benefits, limitations and risks of these models.

Based on the identified gaps, this paper addresses the following research questions:

– Which acoustic (prosodic and vocal) and linguistic markers most consistently relate to key burnout dimensions (e.g., MBI) and under what conditions is this relationship reproducible?

– Which modeling approaches (classical ML vs. DL, including self-supervised speech representations) demonstrate the most potential for in different contexts (different professions, recording channels, languages/code-switching) with limited labeling?

– How can speech, text (ASR/NLP), and emotional analysis (SER) components be theoretically integrated into a conceptual multimodal framework to explore predictive capabilities while preserving interpretability for managerial use?

– What role does organizational context (including interaction metrics and communication structure, ONA) play in explaining and predicting burnout at the team/organization level?

– What ethical and data protection requirements (data minimization, consent, edge approaches) should be incorporated into the design of a speech burnout analytics system to ensure practical applicability in Kazakhstan and comparable regulatory environments?

Research methodology

This study utilizes the PRISMA 2020 statement to identify and analyze existing solutions and to either identify or develop more advanced applications for the modern world using artificial intelligence models.

The Prisma 2020 statement provides a standard for conducting systematic literature reviews, ensuring transparency, reproducibility, and scientific rigor in the synthesis of research findings. According to Page et al. [4], this tool enhances research at every stage, from study proposal to final paper production, establishing clear guidelines to minimize bias and error.

This methodological approach systematically structures the search, selection, and evaluation processes, ensuring reliable results that advance knowledge in this evolving field.

Inclusion and exclusion criteria

To ensure relevance and quality, inclusion and exclusion criteria were predefined that are tailored to the research topic.

Inclusion criteria:

– Publications in peer-reviewed journals and conference proceedings are available in full text.

– Publication period considered papers published from January 2010 to December 2025, to cover the current development of digital phenotyping and self-supervised speech models.

– Studies in which professional burnout or similar topics (emotional exhaustion, depression, chronic stress) are explicit or specific variables are tested; use speech or text data (speech recordings, transcripts, written communications) as a source of features; describe computational models (ML/DL/SSL, SER, NLP) for assessing the state and risk of burnout, stress, or affective disorders.

Exclusion criteria:

– Articles that do not separate burnout from general stress or satisfaction, making it difficult to compare results.

– Papers without access to the full text, as well as studies with insufficient method clarification that make it impossible to evaluate the design, sample, or process for identifying key points.

Sources of information and databases

Databases, including Scopus, PubMed were chosen for the study due to their broad coverage in medicine, psychology, and computational science, making them important resources for identifying strengths of the intersection of quality control and artificial intelligence.

Scopus is chosen because its multidisciplinary, highly curated and continuously updated journals make them particularly suitable for analyzing dynamic fields such as model building. PubMed is known for its suitability for finding research on burnout, stress, and digital phenotyping in clinical and organizational contexts.

Search Strategy

For each database, a combined search string was created using Boolean AND/OR operators, adapted to the syntax of the specific platform. Examples of keywords and their combinations:

For burnout and digital phenotyping:

– "burnout" AND ("digital phenotyping" OR "passive sensing")

– "clinician burnout" AND "speech analysis"

For speech and emotion:

– "speech emotion recognition" AND (stress OR burnout)

– "prosody" AND ("depression" OR "emotional exhaustion")

For models and architectures:

– "self-supervised learning" AND (HuBERT OR WavLM OR "wav2vec 2.0")

– "Whisper ASR" AND ("multilingual" OR "code-switching")

For the organizational context:

– "organizational network analysis" AND burnout

– "call center" AND ("stress detection" OR "voice analytics")

The search was limited to a specified time period, publication type (article, conference paper), and, where possible, subject categories (medicine, psychology, computer science, engineering). Duplicates were automatically removed by the database and then manually checked during export.

Study selection process

The study selection process strictly followed the PRISMA 2020 guidelines [4], and the flow diagram (Figure 1) was generated using the PRISMA2020 Shiny application [5]. An initial search across the Scopus and PubMed databases yielded 697 records. After the automated and manual removal of 62 duplicates, 635 unique records remained for screening. During the title and abstract screening phase, 515 articles were excluded due to a lack of relevance to the predefined intersection

of occupational burnout and speech/text analytics. The remaining 120 full-text articles were rigorously assessed for eligibility. Of these, 103 were excluded for specific methodological reasons: failure to separate burnout from general stress indicators (n = 46), insufficient method clarification or unreplicable design (n = 36), and lack of access to the full text (n = 21). In the end, 17 studies met all inclusion criteria and were synthesized in the final review.

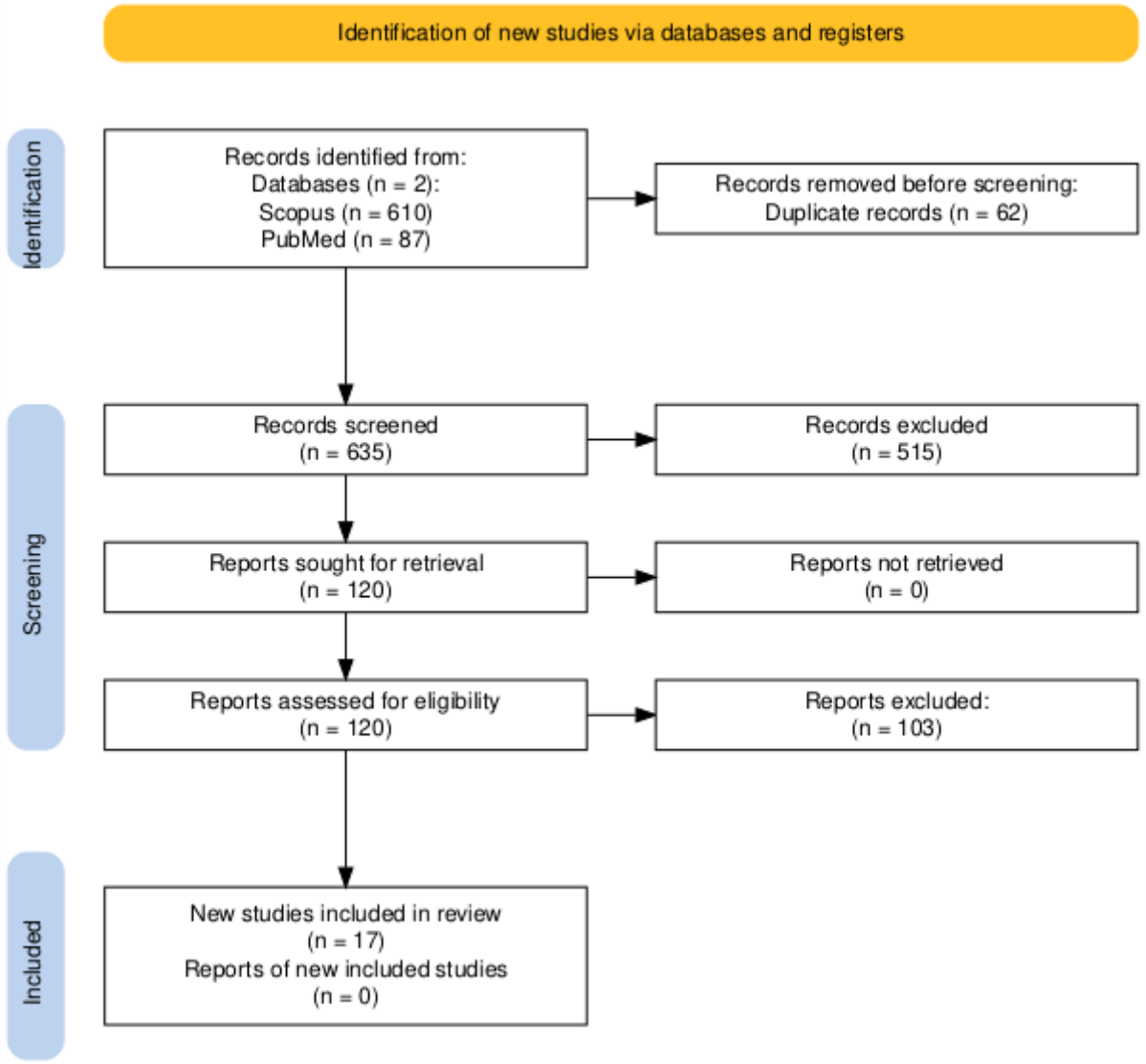

Figure 1. PRISMA 2020 flow diagram for the systematic review, generated via [5]

Data extraction and processing

For each study, a standardized data extraction form was applied where the key sections included:

– General information: authors, year, country, context (medicine, call center, IT company, etc.).

– Target variable: outcome type (MBI burnout, clinical depression, stress scales, composite well-being indices).

– Data type: audio (spontaneous speech, dialogues, interviews, call center); text (transcripts, forums, corporate communications); additional signals (surveys, behavioral metrics, ONA indicators).

– Linguistic and acoustic features: acoustics (F0, jitter, shimmer, speech rate, pauses, spectral features, SSL embeddings); linguistics (pronoun frequency, absolutist index, tonality, neutrality, tense profile).

– Models and architectures: types of algorithms used (SVM, CNN, RNN, transformers, SSL models).
– Quality metrics: accuracy, F1 score, AUC, sensitivity/specificity.

Results of the study

A. Theoretical foundations of burnout and digital phenotyping

1) Burnout Models and the Limitations of Traditional Diagnostics

According to the review of literature, burnout is most often identified via the Maslach Burnout Inventory (MBI) survey, which includes emotional exhaustion, depersonalization (cynicism), and decreased personal accomplishment. Studies show a link between these dimensions and deteriorating mental health, increased turnover, and decreased employee performance. Furthermore, the very logic of these questionnaires assumes that employees are aware of their burnout and are willing to honestly report it.

Criticism of traditional burnout diagnostics has several obvious problems including severely late detection, dependence on cultural norms for emotional expression, and the influence of social desirability. In highly stressed professional cultures where vulnerability is suppressed, questionnaires often only capture a fraction of cases, and burnout only becomes noticeable when the employee's energy is already severely depleted.

2) Digital Phenotyping and Speech as a Biomarker

Digital phenotyping may serve as an alternative which is based on the analysis of data generated by a person during everyday activities, ranging from phone calls to instant messaging [2],[6]. In this sense, speech is significant because it requires cognitive functions and the use of laryngeal muscles, which are also regulated by the hypothalamic-pituitary-adrenal (HPA) axis [3].

Moreover, studies revealed a link between stress hormone levels and vocal characteristics, such as fundamental frequency, its variability, and the appearance of frizz and micro-oscillations. Chronic burnout is associated not only with "high arousal" (e.g., anger or fear), but also with the phenomenon of "emotional blunting," manifested by monotony and decreased variability in speech [3]. Thus, speech can serve as a passive and relatively continuous indicator of the transition from acute stress reactions to chronic exhaustion.

A critical issue in implementing passive observation systems is the "observer effect," whereby employees change their behavior due to the awareness that their speech is being analyzed [7]. According to prior research on participant reactivity (the Hawthorne/observer effect), less obtrusive measurement approaches – such as passive monitoring of existing data streams – can reduce behavior change compared with more active and attention-demanding methods (e.g., frequent self-report surveys) [7]. Therefore, a robust model should prioritize passive data collection to minimize the observer effect and maintain ecological validity.

B. Speech and linguistic markers of burnout

1) Linguistic Markers

Studies that investigated linguistic features of mental state show that not only emotional words but also structural features of language, particularly absolutist thinking, play a significant role [8]. Absolutist thinking is the tendency to describe the world in all-or-nothing terms (always, never, completely, absolutely, should, nothing), which is associated with cognitive rigidity and catastrophizing in depression, anxiety, and suicidal states.

An analysis of texts on mental health forums showed that the frequency of absolutist words is a more accurate marker of psychological disorders than the use of words with negative emotional connotations. In the context of burnout, an employee experiencing depersonalization and cynicism tends to perceive work problems as constant, shifting from "this project is difficult" to "this project is a complete failure."

The "absolutist index" is the ratio of absolutist terms to the total number of words. It serves as a powerful indicator for predictive modeling. Unlike sentiment analysis, which can be misled by sarcasm or professional politeness, absolutist words reveal the speaker's inability to adapt to different

situations. Studies reveal that this linguistic pattern more accurately predicts the degree of psychological distress than negative emotion indices, making it a crucial component of burnout detection systems [8].

Another important marker is the use if pronouns [9]. Since burnout is fundamentally a social phenomenon involving detachment from colleagues and the organization, it is reflected in pronoun use. High-performing teams and active employees typically exhibit a higher frequency of first-person plural pronouns (we, our, us), indicating social integration and shared identity. That is, people experiencing emotional exhaustion and depressive symptoms tend to overuse first-person singular pronouns (such as I, me, my). This pattern is consistent with heightened self-focused attention, reflected in increased use of first-person singular pronouns. In a managerial context, a shift in an employee's writing or speech from "We need to figure this out" to "I'm dealing with it" can serve as an early warning sign of "Depersonalization" burnout (Table 1).

Table 1. Linguistic markers of burnout dimensions

MBI measurement	*Marker*	*Mechanism*
Emotional exhaustion	*Growth of "I" / decline of "we"*	*Self-focusing*
Depersonalization	*Absolutist index*	*Cognitive rigidity*
Reduced achievement	*Fewer future tense verbs*	*Disorientation*
Suppression	*Neutral tonality*	*Emotional dulling*

2) The Neutrality paradox and emotional inertia

Standard sentiment analysis classifiers categorize text as Positive, Negative, or Neutral. In particular, "Neutral" is often interpreted as a baseline or healthy level. However, in stressful professions such as aviation, healthcare, and military operations, "Neutral" can be a misleading indicator.

This predominance of neutrality is largely attributed to professional and cultural norms that discourage the open expression of negative emotions. Personnel in high-stakes environments, such as surgeons or crisis managers, are trained to suppress negativity to maintain operational control. Furthermore, cross-cultural studies on emotional display rules confirm that in certain contexts, suppressing negative emotions and maintaining a neutral facade is a standard social norm rather than an indicator of psychological well-being [10], [11]. Therefore, a predictive model applied to these groups should consider "neutrality" or "reduced emotional variability" as a potential indicator of emotional suppression and burnout, rather than well-being. This phenomenon may indicate an unfavorable state and requires models to detect the inability to switch emotional states in response to context.

3) Acoustic features of chronic stress

Acoustic parameters are more challenging compared to an intentional control of a regular text. While an employee may choose to write a polite email, even the slightest tension in their voice can arise involuntarily.

Under high and prolonged stress, a natural rhythm and a decrease in emotional expressiveness are observed. Veiga et al. [3] show that this fatigue manifests itself most clearly in the flattening of the fundamental frequency (F0). Although temporary stress can cause a sharp rise in pitch, the chronic nature of burnout exhausts the speaker, leading to monotonous speech devoid of emotional range. This may indicate a physiological lack of energy rather a simple reduction of enthusiasm.

Simultaneously, a measurable temporal shift is proceeded during this flat sound. Due to burnout often imposes a heavy cognitive load, the brain's ability to plan sentences and extract words slows. Consequently, this creates a speech pattern marked by significant pauses and a slow speech rate, reflecting the mental fatigue of communicating in a state of exhaustion. At a deeper level, physical fatigue impairs laryngeal control, leading to more intense jitter and shimmer (tiny involuntary

deviations in pitch and volume), which serve as signs that the brain is struggling to maintain its basic functions.

In addition, background noise is a significant challenge in speech emotion recognition (SER) applications. In this sense, "Babble noise" (other people talking) and "office noise" (keyboards, printers, ventilation systems) can seriously degrade the accuracy of acoustic feature extraction [12]. Standard features such as the MFCC are very sensitive to noise. If a model is trained on clean studio data (e.g., RAVDESS) and deployed to a noisy call center, its performance can drop by 20-30%. This requires the use of advanced deep learning architectures capable of separating the speaker's signal from environmental noise, a capability found in modern unsupervised learning models.

Recent studies also confirm the effectiveness of deep architectures for emotion recognition on RAVDESS. This is a hybrid CNN-Transformer model with a cross-attention mechanism achieves 80% accuracy on eight emotions, but the authors note difficulties in recognizing subtle emotions (sadness, disgust), which is consistent with the emotional inertia hypothesis in burnout [13]. Traditional LSTM architectures demonstrate high accuracy on supervised datasets (91.25% on RAVDESS, 98.05% on TESS), but their transferability to natural speech and multilingual scenarios remains poorly tested. The use of transformer models (RoBERTa) in combination with BiLSTM to analyze social media texts confirms the diagnostic value of absolutist constructions and first-person pronouns: the hybrid model achieves 99.4% accuracy in detecting depression. However, such anomalously high accuracy rates in clinical and psychological forecasting strongly suggest overfitting or data leakage. In real-world ecological settings, these metrics are unattainable, highlighting a widespread reproducibility crisis in applied affective computing and the danger of relying on highly controlled, static datasets without cross-domain validation. Importantly, the linguistic patterns of depression and burnout partially overlap, but burnout is additionally characterized by markers of depersonalization and professional cynicism.

C. Models and datasets for burnout recognition from speech

Modern speech-based burnout assessment systems rely on self-supervised speech models (Table 2) that are pre-trained on thousands of hours of unlabeled audio and then subsequently trained for emotion, stress, and burnout tasks [12], [14]. This approach is particularly important in the field of mental health, where labeled data is limited and the quality of predictions depends on the model's ability to capture subtle paralinguistic and linguistic cues.

1) The shift to self-supervision

Early work on emotion and stress recognition employed classical features (e.g., MFCC) and relatively simple classifiers, which led to significant performance degradation when transferred across domains and languages [15]. The emergence of self-supervised learning architectures (wav2vec 2.0, HuBERT, WavLM, and others) has made it possible to train universal speech representations on large corpora and then adapt them to affective states, including burnout risk [12],[14].

The wav2vec 2.0 model served as a foundational breakthrough in this shift, proving the viability of learning from raw audio via a contrastive task [16]. However, while it demonstrates high emotional classification accuracy on clean, laboratory-grade datasets, its performance degrades significantly in noisy, real-world corporate environments, such as overlapping speech and babble noise. This critical limitation in ecological validity necessitated the transition to more advanced clustering and denoising architectures for organizational monitoring [12].

In this sense, HuBERT (Hidden-Unit BERT) learns to predict hidden unit clusters for masked audio fragments, combining ideas from BERT and speech clustering. Using an iterative clustering-prediction scheme, the model achieves state-of-the-art results in speech recognition on LibriSpeech/Libri-Light and serves as a powerful acoustic encoder for emotion and stress tasks [14].

Moreover, WavLM builds on this idea by combining masked speech prediction and denoising, as well as scaling pretraining to 94,000 hours of multi-domain speech. Due to this, WavLM demonstrates high robustness to noise and overlapping speech and achieves top results on the

SUPERB benchmark across a wide range of tasks. This is particularly valuable for burnout-related tasks, as real-world corporate recordings typically contain "office" noise and polyphony [12].

Recent reviews and experiments on audio emotions show that SSL embedding consistently improve emotion classification accuracy, especially when labeled data is scarce, and combine well with lightweight upper-level models [15]. In turn, this creates a convenient path to implementing burnout predictive systems in organizations with limited computing resources.

Whisper, although focused on the ASR task, adds a critical semantic layer, providing reliable transcription in dozens of languages and under code-switching [17]. In the context of burnout, this allows the combination of HuBERT/WavLM acoustic features with linguistic biomarkers (absolutist words, pronouns, polarity) extracted at the text level [8], [9].

Table 2. Comparison of self-supervised speech models

Features	*wav2vec 2.0*	*HuBERT*	*WavLM*	*Whisper*
Type of training	*SSL*	*SSL*	*SSL+denoising*	*Weak sup.*
Primary role	*Acoustics*	*Acoustics*	*Prosody*	*Semantics*
Emotional accuracy	*High (on clean data)*	*High*	*High*	*Moderate*
Noise tolerance	*Low*	*Average*	*Very high*	*High*
Multilingualism	*Limited*	*Limited*	*Moderate*	*Very high*
Code-switching	*Low*	*Low*	*Moderate*	*High*

D. Critical Datasets for Model Training

1) Proposed Architecture: Multimodal Late Fusion

The strategy for training burnout models from speech is typically a multi-staged process. These include pre-training on large general-purpose corpora, training on emotion-specific datasets, and final fine-tuning on specialized clinical or organizational stress/depression samples [15].

In this sense, key emotion-specific datasets include RAVDESS, IEMOCAP, DAIC-WOZ, and EMO-DB, which provide a variety of scenarios ranging from actor's speech to spontaneous dialogues and clinical interviews. These datasets allow models to learn to distinguish between both acute stress responses (increased tone, excitability) and more stable patterns of reduced variability and monotony characteristic of chronic burnout [3], [14]. In the context of Kazakhstan, multilingual and local corpora, such as ISSAI Multilingual Speech, KazE moTTS, and large industrial corpora of Kazakh and Russian speech, play a critical role, allowing HuBERT, WavLM, and Whisper to be adapted to the peculiarities of local prosody, accents, and typical code-switching [18]. Combining global standards with local data creates the basis for robust burnout prediction in the context of real-world office noise, multichannel communication, and culturally specific norms of emotional expression [10].

Regarding this, multimodal integration in related fields confirms the effectiveness of late fusion. For instance, combining raw acoustic embeddings from self-supervised models with semantic features extracted via NLP allows for a much more comprehensive assessment of emotional states than using a single modality [12], [14]. Such hybrid architectures demonstrate that integrating heterogeneous data streams requires not only technical synchronization but also careful tuning of weights and thresholds to ensure the interpretability of the results in an organizational context.

2) Emerging Frameworks: StressSpeak

New framework solutions demonstrate that speech and AI are already being used to monitor stress and burnout, often in scenarios that are closely related to a person's needs. In most cases, these are either voice-based well-being assistants integrated into workflows or specialized platforms for call centers and frontline workers [19].

First, clinically oriented passive monitoring platforms are emerging that analyze voice, text, and behavioral patterns to assess chronic stress and burnout in healthcare workers and other frontline specialists (for example, passive AI systems for assessing stress and burnout using biometric and behavioral data). These solutions use continuous speech collection in daily environments, then apply emotion and stress recognition models to generate risk assessments and recommendations for clinicians or digital therapeutic applications [9]. Secondly, call centers are actively implementing voice analytics platforms that monitor agents' tempo, pitch, pauses, and emotional tone in real time to detect early signs of emotional exhaustion [3]. Examples include commercial "Voice Burnout Detection" systems and products such as Cogito, Insight7, and CallMiner, which alert callers to fatigue-related changes in voice during a call and help managers adjust agent workload and schedules.

Finally, new-generation frameworks are emerging that combine speech, ASR, and large language models, such as StressSpeak, a real-time speech-based system that records audio, transcribes it, analyzes linguistic stress markers using transform LLMs, and immediately provides personalized recommendations. These solutions are moving beyond "bare classification" toward clinical and organizational support ranging from simple stress detection to contextual digital interventions integrated into employee well-being platforms.

All these imply that above-mentioned emerging frameworks confirm the feasibility of the proposed concept, showing that passive speech analysis using SSL models and LLM is already being used to monitor stress and burnout in call centers, telehealth, and among frontline workers. Future implications should consider adding network and organizational context as in architecture with Organizational Network Analysis (ONA).

E. Critical Datasets for Model Training

Training the proposed predictive burnout model requires datasets of two levels: global "standards" for basic emotion/affect learning and local Kazakhstani corpora for linguistic and cultural adaptation. Below are three key datasets for each type.

1) Global Datasets

– RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): 24 actors (12 men, 12 women), 7,356 audio and video recordings with 8 emotions (neutral, calm, happy, sad, angry, fearful, surprise, disgust) at two intensity levels. Studio-quality recordings and validated annotations make RAVDESS ideal for basic training of models to distinguish emotion intensity.

– IEMOCAP (Interactive Emotional Dyadic Motion Capture): Approximately 12 hours of audio and video dialogue, including improvisation and scripts, annotated with categorical emotions and continuum scales (valence, activation, dominance). It is important for training models to recognize spontaneous emotions in conversational speech.

– DAIC-WOZ (Depression and Anxiety Interview Corpus): Multimodal clinical interviews with a virtual agent, aimed at the automatic detection of depression and anxiety disorders. It contains annotations for depression scales.

2) Kazakhstani and multilingual datasets

– ISSAI Multilingual / Kazakh Speech Corpus (KSC): An open corpus of approximately 330 hours of Kazakh speech collected from speakers of different regions and ages. It provides the model with robustness to accents and pronunciation variability.

– KazEmoTTS (Kazakh Emotional Text-to-Speech): 74.85 hours of high-quality Kazakh speech (54,760 audio-text pairs) in 6 emotional categories including neutral, angry, happy, sad, scared, surprised which is recorded by three professional speakers.

– Kazakh Speech Corpus 2 (KSC2) is the first ever industrial-scale open corpus that encompasses approximately 1,200 hours of transcribed speech (600k utterances) from TV, radio, parliament, podcasts, and more, including cases of Kazakh-Russian code-switching.

F. Organizational Network Analysis

Organizational Network Analysis (ONA) perceives an organization as a network of connections between people and allows for the measurement of not only the formal structure but also the actual

flows of influence, support, and workload. Unlike individually focused burnout metrics, ONA identifies central connectors with high network overload and the risk of burnout contagion, as well as peripheral nodes that combine low network engagement with an increased risk of isolation and depression [20].

For the predictive burnout model, ONA serves as a contextual "grounding." Particularly, it changes in voice and language are interpreted differently depending on the employee's position in the network and their role in communication flows. In turn, this allows distinguishing situational acute stress and chronic exhaustion not only at individual but also at the system level. This integrated consideration of individual speech biomarkers and network metrics makes the model more useful for managerial decisions from workload redistribution to targeted team support programs.

G. Cultural and Legal Localization: The Kazakhstan Context

The Kazakhstani context fundamentally influences the design of an ethically and legally sound speech-based burnout prediction system. Unlike typical Western settings, the model simultaneously proceeds with the unique features of emotional expression of the post-Soviet culture and a hybrid legal regime of data protection. [10], [21]

1) Cultural Rules of Emotional Expression

Research on post-Soviet cultures shows a consistent norm of restraint in negative emotions and a more cautious attitude toward displaying joy to strangers, compared to the United States and Western Europe [10]. In such an environment, a "neutral" manner of speech and facial expression may reflect not psychological well-being, but a social norm of emotional restraint, which requires local calibration of baseline valence and arousal levels during model training [11].

For Kazakhstan, the phenomenon of everyday code-switching between Kazakh, Russian, and English in professional communication is also important. Language switching is used as a tool for identity and efficiency, and not simply as a sign of proficiency deficit [18]. This means that linguistic markers of burnout (absolutist words, pronouns, "neutrality") must be extracted from multi-code speech, and the ASR component must reliably recognize mixed utterances; otherwise, some semantic features will simply be lost.

2) Legal Regulation: Law No. 94-V and AIFC

The main law on personal data in Kazakhstan is Law No. 94-V "On Personal Data and Their Protection," which regulates the collection, processing, and storage of personal and biometric data, including voice. The law requires a specific, clearly defined processing purpose, minimization of collected data, and obtaining the consent of the subject. When working with biometrics (including voice parameters), this usually requires written consent with a separate statement of the purposes and scope of processing.

At the same time, labor laws and practices allow for the processing of employee data for law enforcement and labor organization purposes, provided that the relevant conditions are outlined in the employment contract and the employer's internal regulations. As a result, the system is often more flexible than in the EU, meaning that with a transparent policy, clear information, and employee consent, the employer can implement monitoring systems, including speech analysis, as long as the principles of lawfulness, reasonableness, and security of processing are maintained. [21]

Unlike this, the Astana International Financial Center (AIFC) has a separate system, which has adopted its own Data Protection Regulations, largely modeled on the GDPR (AIFC, 2017). For AIFC member companies, the requirements are more similar to those in Europe, which include mandatory legal foundations (consent/legitimate interest), the principle of "data minimization," DPIAs for high-risk processing, and enhanced requirements for working with "sensitive data," including voice biometrics.

From a cultural perspective, the model should take into account a higher "threshold" for overt emotional expression and perceive persistent neutrality and a reduced range of emotions not as the norm, but as a possible marker of professional masking or exhaustion, especially in high-stake positions. From a linguistic perspective, the architecture must support robust multilingual ASR and

code-switching analysis to accurately extract linguistic burnout indices from the real and particularly multilingual working environments as in Kazakhstan.

From a legal perspective, the model design in Kazakhstan must include explicit written consent from employees, considering the data use not for disciplinary measures but exclusively for well-being purposes. Also, it needs local or edge processing, when possible, and minimizing the storage of raw voice recordings. For companies operating within the AIFC, GDPR-like practices are required, such as conducting DPIAs, ensuring strict purpose limitation, and documenting the legal basis for data processing. In practice, this necessitates a two-tier compliance policy: one aligned with Kazakhstan's general regulatory regime and another tailored specifically tailored to the AIFC jurisdiction.

Discussion

A. Technological Limitations and Computational Barriers

This review demonstrates that while the transition from episodic questionnaires to continuous digital speech phenotyping offers significant potential, several technical hurdles remain unresolved. Modern self-supervised learning (SSL) architectures, such as HuBERT and WavLM, have achieved state-of-the-art results in speech emotion recognition (SER) by training on massive unlabeled datasets. However, these models require substantial computational resources that may exceed the capacity of standard office hardware, creating a barrier to local or "edge" implementation. Furthermore, there is a significant performance gap when moving from studio-recorded datasets like RAVDESS to real-world corporate environments. Background "babble noise" and office equipment sounds can degrade the accuracy of acoustic feature extraction by as much as 30%, necessitating more robust denoising layers that are still in development. Finally, the "black box" nature of deep learning models complicates the interpretability of results for managerial use. Without transparent metrics explaining why a high-risk score was generated, managers may lack the confidence to initiate supportive interventions, leading to a reliance on algorithmic verdicts rather than professional judgment.

B. Cultural Ambiguity and the Neutrality Paradox

The application of speech analytics in the Kazakhstani context reveals a unique "neutrality paradox" that challenges standard emotion classifiers. In post-Soviet professional cultures, social norms often dictate a high level of emotional restraint and caution in expressing joy or frustration to outsiders. Consequently, a "neutral" vocal tone—often interpreted by Western-trained algorithms as a sign of stability—may actually serve as a marker of emotional blunting or "masking" associated with deep burnout. This cultural specificity requires local calibration of baseline arousal and valence levels to avoid high rates of false negatives. Additionally, the prevalence of code-switching between Kazakh, Russian, and English in Kazakhstan's workplaces complicates linguistic analysis. If the automatic speech recognition (ASR) component fails to capture mixed-language utterances accurately, critical semantic markers such as the "absolutist index" or shifts in pronoun use may be lost, rendering the linguistic profile incomplete.

C. Ethical Risks and the Hawthorne Effect

Implementing passive speech monitoring systems introduces profound ethical risks, primarily the "observer effect" or Hawthorne effect. When employees are aware that their vocal intonations and linguistic choices are being monitored for signs of burnout, they may consciously or unconsciously alter their speech to appear more productive or emotionally stable. This reactivity can undermine the ecological validity of the data and potentially increase the very stress the system is designed to detect. From a legal perspective, organizations in Kazakhstan must navigate a complex two-tier regulatory regime involving General Law No. 94-V and the more stringent, GDPR-like AIFC Data Protection Regulations. To ensure compliance and maintain employee trust, systems must strictly adhere to the principle of "data minimization" by processing voice recordings locally and storing only anonymized embeddings. There is also a persistent risk that these tools could be repurposed for disciplinary control rather than wellness, necessitating clear firewalls between psychological risk data and performance management metrics.

D. Integration with Organizational Network Analysis

A key advancement identified in this review is the shift from viewing burnout as an individual psychological failure to a systemic organizational issue. By integrating speech biomarkers with Organizational Network Analysis (ONA), managers can interpret vocal changes within the context of an employee's social and functional role. For example, a decline in prosodic variability in a "central connector" might indicate burnout contagion across a team, whereas similar markers in a peripheral node might suggest social isolation. This integrated approach allows for more targeted interventions, such as workload redistribution or enhanced team support, rather than placing the burden of recovery solely on the individual employee. However, the use of ONA also increases the "digital vulnerability" of key employees, as their high visibility in the network may make them targets for hidden discrimination if their risk scores become known.

Conclusion

This systematic review presents a structured synthesis of research on predicting professional burnout based on speech data analysis, combining psychological models of burnout with computational approaches such as machine learning and digital phenotyping. A literature review conducted in accordance with the PRISMA 2020 guidelines revealed consistent acoustic and linguistic patterns associated with key burnout dimensions, including pitch reduction, reduced prosodic variability, and the phenomenon of emotional inertia at the acoustic level, as well as an increase in the proportion of absolutist constructions and first-person singular pronouns in the speech of individuals with high emotional exhaustion. However, the review demonstrated high methodological variability in data collection protocols, operationalization of target constructs, and choice of model architectures, which complicates direct comparison of results and limits the replicability of findings across different organizational contexts.

The transition to self-supervised speech representations such as HuBERT, WavLM, and wav2vec 2.0 opens new possibilities for building robust predictive models with limited labeling and domain shifts, including multilingual scenarios and code-switching conditions typical of post-Soviet and developing regions. Hybrid architectures combining acoustic embeddings, textual representations (ASR → NLP), and emotional analysis components (SER) via late fusion demonstrate advantages in accuracy and modularity, but their interpretability for management decisions and clinical validity remain insufficiently explored. A critical barrier to practical application is the lack of specialized datasets that include speech in authentic work settings with standardized labeling for burnout measures and contextual metadata about the profession, team structure, and organizational stressors.

Integrating organizational context through Organizational Network Analysis expands the explanatory potential of models, allowing them to account for structural characteristics of communication, the social contagion of burnout, and the effects of isolation or role overload. Such multilevel approaches, combining individual speech markers with network metrics, are still in the early stages of development but represent a promising avenue for targeted interventions at the team and department levels. In practical terms, the obtained results form the basis for piloting early risk detection systems in controlled settings such as corporate call centers, telemedicine, and frontline services with a mandatory focus on wellness applications rather than performance monitoring to minimize the risk of stigmatization and maintain employee trust.

Geographical research in speech analytics for burnout is concentrated primarily in North America and Europe, with significant disparities in access to computing resources, high-quality datasets, and regulatory frameworks for biometric data in developing regions. Localizing such systems for specific jurisdictions, particularly in Kazakhstan, with its data protection policies (Law 94-V) and optional AIFC regime, similar to the GDPR requires the implementation of data minimization protocols, explicit consent, edge processing, and DPIA for speech biomarkers. Cultural specificities, including restraint in emotional expression and the prevalence of code-switching must also be reflected in the selection of pre-training corpora and model tuning, highlighting the need for international collaboration and open exchange of methodologies.

In conclusion, the application of computational speech analysis methods to burnout prediction holds transformative potential for organizational health and risk management, but large-scale adoption depends on progress in several key areas. Future research should prioritize the development and publication of longitudinal datasets with natural speech, standardized MBI labeling, and contextual variables; the development of hybrid architectures with explicit controls for biases (linguistic, cultural, and demographic); the calibration of thresholds for clinically significant changes based on prospective validation studies; and the creation of ethical frameworks and regulatory standards that balance the potential benefits of early detection with the protection of workers' rights and dignity. Only by meeting these conditions will speech analytics for burnout realize its potential in real-world organizational scenarios.

Acknowledgement

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510633).

References

[1] Maslach, Christina & Jackson, Susan & Leiter, Michael. (1997). The Maslach Burnout Inventory Manual. Evaluating Stress: A Book of Resources (pp.191-218).

[2] Onnela, J. P., & Rauch, S. L. (2016). A Call to Expand the Scope of Digital Phenotyping. Journal of Medical Internet Research.

[3] Veiga DL, Almeida TM, Uchida RR, Cordeiro Q. The Fundamental Frequency of Voice as a Potential Stress Biomarker: A Systematic Review and Meta-Analysis. Stress Health. 2025 Oct;41(5):e70112.

[4] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, D. Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ.

[5] Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews, 18, e1230.

[6] Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, Schoeller F, Mouchabac S, Adrien V. Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. J Med Internet Res.

[7] Goodwin MA, Stange KC, Zyzanski SJ, Crabtree BF, Borawski EA, Flocke SA. The Hawthorne Effect in Direct Observation Research with Physicians and Patients. Eval Clin Pract. Author manuscript; available in PMC: 2018 Dec 1.

[8] Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. Clinical Psychological Science.

[9] Teo, C., et al. (2023). The language of healthcare worker emotional exhaustion: A linguistic analysis. PubMed Central.

[10] Sheldon, K. M., et al. (2017). Russians Inhibit the Expression of Happiness to Strangers: Testing a Display Rule Model. Journal of Cross-Cultural Psychology.

[11] A. Pankratova, Evgeny N. Osin (2020). Circumplex model of emotional display rules: a cross-cultural study. Psikhologicheskii zhurnal.

[12] Chen, S., et al. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. Microsoft Research.

[13] Zhongliang Wei, Chang Ge, Chang Su, Ruofan Chen, Jing Sun (2025). A Deep Learning Model for Speech Emotion Recognition on RAVDESS Dataset. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 5, 2025.

[14] Hsu, W. N., et al. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.

[15] Peranut Nimitsurachat, Peter Washington. Self-Supervised Learning for Audio-Based Emotion Recognition.

[16] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[17] Radford, A., et al. (2022). Robust Speech Recognition via Large-Scale Weak Supervision (Whisper). OpenAI.

[18] Zharkynbekova, S., Shakhputova, Z., Anichshenko, O., & Agabekova, Z. (2025). The Speech Behaviour of Kazakhstani Youth in the Context of Interethnic Communication. Journalism and Media, 6 (1).

[19] Kumar, Gowtham & Mathew, John. (2025). Speech Emotion AI for Mental Health Monitoring in Call Centers. International Journal Of Advance Research And Innovative Ideas In Education.

[20] Decker-Tonnesen PL, Chesak SS, Walker LE, Kohler K, Phelan S, Gunnels MS, Saliba KL and Bhagra A (2025). Role of organizational network analyses to advance workforce inclusion and belonging: a scoping literature review. Frontiers in Psychology.

[21] Republic of Kazakhstan. Law No. 94-V: On Personal Data and their Protection. Legislationline.