SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(1), 528.
https://doi.org/10.62617/mcb528

Article

A planning decision support model integrating bioinformatics and occupational health data with an emphasis on biomechanics

Jing Li[1], Wei Liu[2,*], Feifei Chen[3]

[1] Shijiazhuang Institute of Railway Technology, Shijiazhuang 050041, Hebei, China
[2] Hebei Chemical & Pharmaceutical College, Shijiazhuang 050026, Hebei, China
[3] Shijiazhuang Institute of Railway Technology, Shijiazhuang 050041, Hebei, China
*** Corresponding author**: Wei Liu, yueyue1563018@126.com

CITATION

Li J, Liu W, Chen F. A planning decision support model integrating bioinformatics and occupational health data with an emphasis on biomechanics. Molecular & Cellular Biomechanics. 2025; 22(1): 528. https://doi.org/10.62617/mcb528

ARTICLE INFO

Received: 12 October 2024
Accepted: 11 November 2024
Available online: 2 January 2025

Abstract: In today's rapidly evolving workplace environments, the integration of bioinformatics with occupational health data presents a unique opportunity to enhance employee well-being and optimize workplace safety, especially from the perspective of biomechanics. Existing systems often fail to account for individual genetic factors and the biomechanical aspects of the work environment when assessing occupational health risks, resulting in an increase in workplace-related health problems and less effective health treatments. The primary objective of this study is to develop a planning decision support model that integrates bioinformatics and occupational health data to recognize health risks and generate tailored interventions for employees. Incorporating biomechanics, we explore the impact of physical factors such as workstation ergonomics, repetitive motion patterns, and force exertion levels in the work environment on employee health, and analyze their relationship with genetic predispositions. For example, we study how specific genetic traits may interact with biomechanical stressors to increase the likelihood of musculoskeletal disorders. Initially, study data were collected from diverse sources, including bioinformatics databases and occupational health records, ensuring a comprehensive dataset for effective model training and validation. Data cleaning and Z-score normalization were used in the data preparation stage. Feature extraction was performed using Linear Discriminate Analysis (LDA) to reduce dimensionality from preprocessed data. Data fusion was accomplished by sharing information between bioinformatics and occupational health datasets, enabling a more comprehensive decision support model. The study proposed a Dynamic Bacterial Foraging fine-tuned Efficient Adaptive Boosting (DBF-EAdaBoost) method that integrates dynamic bacterial foraging optimization with adaptive boosting techniques to significantly enhance classification performance in bioinformatics and occupational health data analysis. The proposed algorithms offer high accuracy (0.93), precision (0.987), brier score (0.100), AUC (0.92), and log loss (0.314) in forecasting potential health issues based on workplace exposures, biomechanical factors, and genetic predispositions. To enhance the practicality of the research, a more detailed explanation of the implementation process and advantages of the proposed DBF-EAdaBoost algorithm is provided. Consider including real-world case studies to demonstrate the model's application and the actual effectiveness of health interventions in real workplace environments. For instance, we can present a case where the model was applied in a manufacturing plant to predict and prevent musculoskeletal disorders among workers by analyzing their biomechanical workloads and genetic profiles, and implementing appropriate ergonomic interventions. The planning decision support model serves as a significant tool for public health officials, policymakers, and occupational health professionals, promoting data-driven decisions that enhance health outcomes.

Keywords: occupational health; planning decision support model; bioinformatics; dynamic bacterial foraging fine-tuned efficient adaptive boosting (DBF-EAdaBoost); biomechanics

1. Introduction

Advances in healthcare are providing new options for developing and implementing patient-centered care (PCC) models across medical practices of all sizes [1]. Improved communication and ecosystems help the patients, which is very important for success. PCC bridges the gap between patients, their loved ones, and their medical conditions. It focuses on communication between healthcare professionals, patients, or caregivers [2]. It defines patient-centered care as "respectful and responsive to specific needs and desires of patients, with patient values guiding all clinical decisions [3]." PCC aims to improve care, and well-being, resolve disparities, provide value for money, promote individual freedom, and prevent abandonment. PCC attempts to provide tailored care by increasing patient involvement. Effective medical care involves bringing together patients with physicians on a single platform to monitor their health by analyzing daily activities [4]. It promotes collaboration between healthcare stakeholders, provides adequate services, informs decision-making, and optimizes the use of resources. Smart healthcare utilizes the Internet and portable Internet technology to constantly obtain data, link healthcare stakeholders, and intelligently manage and respond to medical requirements [5]. Personalized data in medicine offers the potential for treatment and diagnosis at the individual patient level. Computational models help identify illness causes and variables despite large and heterogeneous datasets [6]. Additionally, they allow for specific treatment plans that depend on crucial medical issues. Computational models can translate in-vitro, preclinical, and clinical outcomes, including uncertainty, into diagnostic or prediction expressions.

Bioinformatics integrates the disciplines of biology, physics, chemistry, statistics, and computer science to address challenging biological phenomena. Informatics is a rapidly developing and versatile scientific discipline [7]. Bioinformatics aims to preserve, analyze, and retrieve information on creatures to better understand their dynamics. Bioinformatics is built on data that has been supplied and generated. The human body's cells use deoxyribonucleic acid (DNA) to interpret data and anticipate illness risk. Genetics plays a crucial role in medical practice, allowing for the accurate identification of diverse disorders [8]. It improves disease prognosis and helps patients choose the best treatment options. The ability to analyze the human genome at several levels, including chromosomal and single-base alterations, adds to its current potential. The availability of massive digital medical datasets has enabled the application of informatics to medical care and research, opening up new avenues for discovery and exploration. Informatics aims to create effective strategies for processing information using technology [9]. Informatics is widely used in healthcare, from research to service delivery, with several specializations including bioinformatics, medical informatics, and biomedical informatics.

Occupation and work status are key socioeconomic determinants of health. Work impacts money, relationships, and access to education, housing, food, and healthcare. Work-related information is often intertwined with other health determinants like race, ethnicity, gender, and citizenship status, making it difficult to guide clinical decision-making and population health activities [10]. It is also

understudied as an essential aspect of health. Many health data collection systems, such as certificates of death, cancer databases, and provincial health department case reports, require employment and industry information. These systems include paper, digital distribution of paper, web-based forms, and upcoming compatible information technology (IT) systems.

The study's purpose is to develop a decision-support tool that uses bioinformatics with occupational health data to detect workplace health risks. Its goal is to provide data-driven insights into wellness interventions and policies to enable individualized health planning, increase workplace security, and optimize decision-making. The suggested model, DBF-EAdaBoost, combines bioinformatics and health data to improve prediction accuracy for health hazards. Using this paradigm, the project aims to provide more accurate and effective health interventions, eventually building a safer and healthier workplace environment through customized, data-driven methods.

Key contributions of the study

- The work uses a variation of databases, with bioinformatics databases and occupational health records, to ensure a huge dataset for efficient model training and validation.
- The study uses data cleaning and *Z*-score normalization throughout the data preparation phase to improve data quality.
- Linear Discriminant Analysis (LDA) is used to extract features, reduce dimensionality, and preserve class discriminant qualities.
- The Dynamic Bacterial Foraging fine-tuned Efficient Adaptive Boosting (DBF-EAdaBoost) approach enhances classification performance when assessing health data.
- A comparison analysis is maintained to assess the presentation of the proposed DBF-EAdaBoost approach versus existing algorithms, demonstrating its superior precision and effectiveness in health risk prediction.
- The decision-support approach improves health outcomes by combining bioinformatics with occupational health data. This allows public health officials and politicians to make data-driven decisions that increase occupational health risk assessments and interventions.

The paper is divided into eight phases: Phase 1 offerings the subject matter, Phase 2 examines related works, Phase 3 describes the methodology used, Phase 4 presents the decision support model, phase 5 is the experimental setup for the model, Phase 6 discusses the findings of the study, Phase 7 provides the discussion, and Phase 8 summarizes the study's conclusion and future directions.

2. Related work

Study [11] analyzed two Bayesian models, Unigram (UNB) and Bigram (BNB) Naïve Bayes, to autocode severe injury narratives from OSHA data. The dataset included reports of injuries from January 2015 through February 2021. Data preprocessing entailed selecting cases according to model agreement and forecast probability criteria. The results show that the UNB model has a sensitivity of

75.21%, slightly higher than BNB's 75.17%. The combined filtering strategy increased sensitivity to 88.17%, identifying 31% for human evaluation and achieving an F1-score of 55% for top predictions.

Paper [12] described an intelligent clinical decision support system for breast cancer prevention that addresses variability in diagnostic process interpretation. By combining trained systems with fuzzy logic, exploratory analysis of factors, augmenting data, and algorithmic classification, the system examined patients' medical data and created a cancer risk alert level. Initial performance testing on a 130-case database from the University of Wisconsin-Madison yielded ROC curve areas of 0.95 to 0.97, indicating substantial diagnostic and preventative potential for clinical use.

The present research [13] looked at the difficulties that mental health practitioners encounter when identifying diseases such as depression and anxiety, which frequently require sophisticated and time-consuming assessments. To improve effectiveness and precision, a decision support system (DSS) based on advanced analytics and AI was built. The DSS used the Networked Probabilistic Pattern Recognition (NEPAR) algorithm to shorten the evaluation procedure, which requires only 28 targeted questions from participants. Machine learning models undergo training to predict mental diseases with an 89% accuracy rate. This method increased participation rates and improved clinical decision-making for psychological practitioners.

The goal developed by [14] examined the intelli-Omics was to provide an adaptable decision support system for multi-omics analysis of data, allowing personalized medicine through rapid data integration, research, and analytics. Data was collected using high-throughput technology and analyzed in Hadoop, with Apache Hive translating it into a knowledge base. Apache Spark &MLlib handle the extraction of features and ML, respectively. The method facilitated clinical decision-making, notably in non-small cell cancers of the lung therapy, by providing personalized reports. While scalable and configurable, it necessitated technical knowledge and large computational resources.

Author [15] addressed the contemporary advances in workplace technologies that improve worker health, security, and productivity. It used data from various mobile devices and connected employee solutions to track workplace dangers and injuries. Preprocessing ensured consistency by integrating several data sources, whereas feature extraction found crucial safety and health measures. The methodologies entail examining the functionality of these advances and their real-world implementations. The results show that wearables were useful for monitoring ergonomic practices and tiredness, as well as predictive analytics that improve making decisions and risk management in workplace situations.

According to the author of [16], the Human Health Exposure Analysis Resource (HHEAR) improved knowledge about exposure to the environment and its effects on human health across the life cycle. It solved difficulties in including evaluations of exposure to research on health, such as restricted researcher expertise and laboratory access. HHEAR promoted collaborative research by providing free scientific data analysis and processing expertise. Its capacity to link biological specimens and environmental samples allows for more advanced analyses that link exposures to

health consequences, which benefits the entire scientific community.

Author [17] investigated Convolutional Neural Networks (CNN) for assessing complex medical information in spine surgery. It focused on gathering data from genetic, radiological, and therapeutic sources. Preprocessing entailed transforming non-imaging information into images for CNN input, allowing feature extraction. The strategy integrated multi-input data using hybrid deep learning models. The findings indicate promising advances in the recognition of patterns, with benefits such as improved decision-making and individualized patient care, while obstacles include the complexity of data and a requirement for multisensory interaction.

Article [18] emphasized the transformation in medicine from restricted clinical data to huge, diversified data streams, which require new tools for physicians. It outlined the creation of a medical bioinformatics program to assist with the shift to a teaching health system by offering training and demonstrating informatics functions. Good criticism of educational programs indicated increased involvement. The result underlined the significance of doctors assessing data quality and usefulness, as well as comprehending AI along with predictive analytics, to adapt to the massive data era, as demonstrated by initiatives such as Wake Forest's program.

The study aimed [19] to enhance the forecasts of unplanned hospital readmissions by integrating the "Unplanned Readmission Model version 1" with the Epic electronic health record. Data was obtained over two years, demonstrating predictive skills with AUC/C statistics for all patients and general medical patients. The model's positive value for prediction ranged from 0.217 to 0.248. The paper also described how to evaluate trends and solutions for reducing readmissions caused by predictive scores.

Based on [20] examined Turkey's National Health Information System (NHIS-T) in terms of genetic data interoperability for improved molecular diagnosis and individualized treatment. It underlined the relevance of data security and privacy rules by contrasting Turkey's "Law on the Protection of Personal Data" to worldwide norms. The authors address the importance of a national standard database and IT infrastructure for integrating genetic data with health information. Established terminology, government direction, and clear guidelines for moral structures and public interaction were all critical success elements.

3. Methodology

The process involves collecting a variety of data in occupational health records from bioinformatics databases. Data are first processed using *Z*-score normalization to standardize the values. Feature extraction uses linear discriminant analysis (LDA) to reduce dimensionality, ensuring that the system focuses on relevant attributes for advanced health risk assessment and personalized treatment, as shown in **Figure 1**.

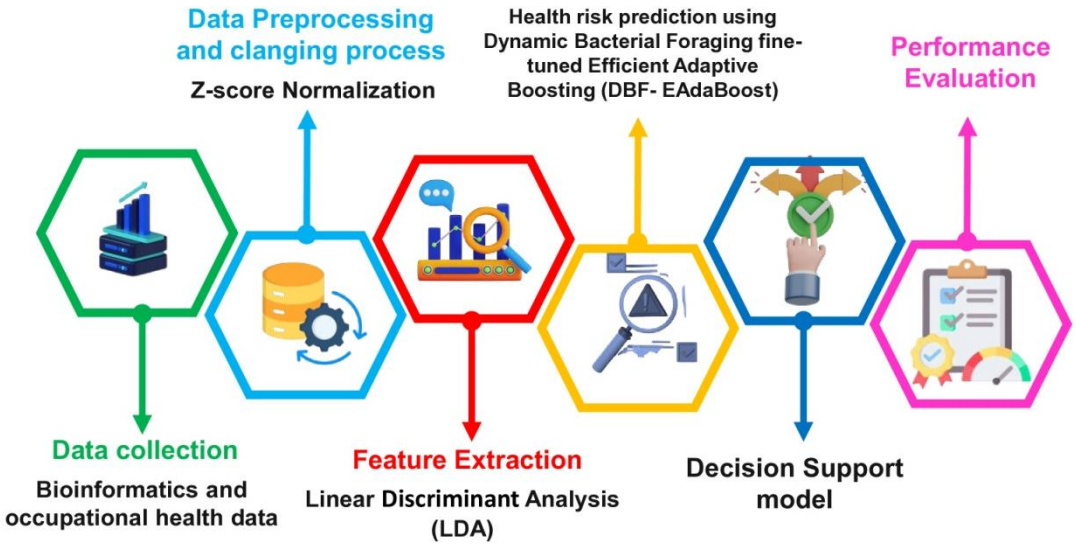

Figure 1. Basic concept of the proposed research framework.

3.1. Data collection

Table 1. Dataset description of the employee health and occupational data.

Employee ID	Age	Gender	Employment duration (Years)	BMI	Blood pressure (mmHg)	Cholesterol (mg/dl)	Fasting blood sugar (mg/dl)	SNP data	Job stress level (1–10)	Health outcome
01	37	M	10	24.5	190	220	85	rs123456 A/G	8	None
02	28	M	5	22.1	180	210	98	rs789012 C/T	5	Mild Respiratory Issues
03	44	F	7	30.1	210	180	110	rs345678 T/C	6	Back Pain
04	48	M	15	28.5	200	195	95	rs456789 G/A	7	None
05	35	F	12	27.2	170	215	120	rs567890 C/G	8	Fatigue
06	35	F	15	29.6	200	186	80	rs678901 T/G	6	Respiratory Issues
07	26	M	5	24.3	180	208	112	rs890123 C/A	5	None
...	...	...	...	...	...	...	...	...	...	...
n	n	n	n	n	n	n	n	n	n	n

The database includes health records for 500 employees who receive monthly assessments to identify risk factors within the organization. This collection emphasizes bioinformatics and occupational health data, including data on various health indicators, genetic data, and occupational exposures. The data collection seeks to reveal potential health problems by studying the relationship between individual health considerations and the work environment. This approach ensures proactive planning to address health risks within the facility, as displayed in **Table 1**. It captures demographic, genetic (SNP data), employment duration, and occupational details like job stress, alongside health metrics such as BMI, blood pressure, and

cholesterol. This data aims to reveal correlations between health outcomes and workplace exposures, supporting proactive health management within the organization.

3.2. Data preprocessing using *Z*-score normalization

Z-score standardization is used to translate healthcare-related information into a single scale, making it easier to compare diverse data points for 500 employees. The *Z*-score equation using raw data value Y, is used as shown in Equation (1) to normalize attributes, such as age, gender, blood groups, glucose level, and test outcomes, rendering them easier to analyze and understand across variables between models.

$$w_{ji} = Y(w_{ji}) = \frac{w_{ji} - \overline{w_i}}{\sigma_i} \tag{1}$$

Deviation from the *j*-th feature will be corrected using *Z*-score standardization. The consequent variable will have a median of 0 and a variance of 1, but its location and scale will be lost. This strategy is only applicable to global standardization; hence, it can't be efficiently employed for particular healthcare subgroups or clusters. Thus, while *Z*-score standardization is good for general analysis, it could not capture variances within specific patient categories.

3.3. Feature extraction using linear discriminate analysis (LDA)

After the *Z*-score normalization, Linear Discriminant Analysis (LDA) was used to extract features. LDA assisted in transforming the dataset into a lower-dimensional space, increasing the separation of classes by minimizing within-class variation and decreasing between-class variance. This technique ensured that critical data structures were preserved while refining the characteristics for improved accuracy in classification in the evaluation. We use the following Equations (2)–(9).

$$Tx = \sum_{w \in C_j}^{d} Tx_j \tag{2}$$

$$Tx_j = (w_j - \mu_j)(w_j - \mu_j)^S \tag{3}$$

$$\mu = \frac{1}{M} \sum_{j=1}^{M} w_j \tag{4}$$

$$Ta = \sum_{j=1}^{d} M_j Ta_j \tag{5}$$

$$Ta_j = (n_j - n)^2 \tag{6}$$

$$(n_j - n)^2 = X^S(\mu_j - \mu)(\mu_j - \mu)^S X \tag{7}$$

$$Ta = \sum_{j=1}^{m} M_j(n_j - n)(n_j - n)^S \tag{8}$$

$$N_j = \frac{1}{m} \sum_{w \epsilon C}^{m} W_l \tag{9}$$

Ta Indicates the between-class variance with Tx, w representing the within-class variance, d as the total number of unique courses, as the input value, M as the number of samples in a given class l. To compute the mean (N) of each input (W) for each class (l), divide the average of numbers by the total amount of samples. The eigenvalue of the transformation matrix X is employed for the extraction of features in LDA to ensure optimal class separation. The feature extraction strategy based on the data collection improves the model's capacity to distinguish across classes, increasing classification accuracy and lowering dimensionality while maintaining critical information.

3.4. Health risk prediction using dynamic bacterial foraging fine-tuned efficient adaptive boosting (DBF-EAdaBoost)

Dynamic Bacterial Foraging Fine-Tuned Efficient Adaptive Boosting (DBF-EAdaBoost) is an optimization method that improves classification performance in the designed decision-support model. It uses bacterial foraging principles to dynamically adjust sample weights based on classification incorrect rates, thereby increasing the accuracy of the health risk prediction and the effortless integration of occupational health information leads to more informed decision making. In this phase, this optimization framework is used to fine-tune the model, demonstrating the effectiveness of combining adaptive amplification with naturally occurring mechanisms to generate better health outcomes of the prediction.

3.4.1. Efficient adaptive boosting (EAdaBoost)

In this study, the Efficient AdaBoost algorithm is used to efficiently categorize health hazards based on bioinformatics and occupational health datasets from 500 employees. These databases include genetic information, health measures, including occupational exposure data. AdaBoost improves model performance by iteratively modifying sample weights, focusing on examples that are difficult to classify. By integrating weak learners, the predictive capability of the planned decision support model is improved, allowing for more accurate evaluations of health hazards and better decision-making for employee well-being and workplace safety. The database is represented as $T = \{(w_1, z_1), (w_2, z_2) \dots (w_M, z_M)\}$, which refers to the total amount of datasets used in the training phase. This structured strategy ensures that the algorithm efficiently uses the data acquired from Linear Discriminant Analysis (LDA) to improve accuracy in classification when forecasting potential hazards based on working conditions with genetics. Initialize the weights of all the samples in the initial training data (vector D). Every specimen has an identical gravity (1 over N). The training was carried out using an inadequate learning algorithm called h1. After training, the mistake rate was computed using Equation (10). M_{err}Represents

the amount of erroneously classified observations.

$$\varepsilon = \frac{M_{err}}{M} \tag{10}$$

Determine the relative importance of the weak learning method. The amount of weight of the weak learner technique is determined by the mistake rate and expressed in the vectorα using Equation (11).

$$\alpha = \frac{1}{2} In\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{11}$$

Update the weight of each sample. If this is another instance of misclassification, Equation (9) will be triggered. In some circumstances, the conventional weight updating procedure will continue to be applied. Following t-round learning, the weight and results of each weak predictor are determined. Equation (12) shows the method's outcome.

$$G(W) = sigm\left(\sum_{j=1}^{s} \alpha_j g_j(W)\right) \tag{12}$$

By initializing equal weights for all samples and employing a weak learning algorithm, we iteratively update sample weights based on misclassification rates, allowing the model to focus on challenging instances.

3.4.2. DBF

The DBF optimize technique is applied within the planning decision support model to enhance the identification and management of workplace health concerns. The DBF method employs a bio-inspired search strategy with a complex framework that enables iterative optimization, resulting in optimal or near-optimal solutions. The improved version of DBF focuses on increasing optimization efficiency while preserving the cooperative and competitive dynamics inherent in the three-layer Bacterial Foraging Optimization (BFO) structure. By integrating these advanced optimization techniques, data-driven insights inform wellness interventions and policies, ultimately promoting better health outcomes in the workplace.

Step 1: The BFO variables (a maximum number of movement times M_d, reproductive times M_{qf}, elimination-dispersal durations M_{fc}, population size N, and number of swimming times M_t) were established.

Step 2: Bacterial position is initialized using Equation (13), and the initial fitness value is specified as W, where Rand is a random number that is evenly distributed between 0 and 1.

$$W = w_{min} + Rand(w_{max} - w_{min}) \tag{13}$$

Step 3: Consists of the elimination-dispersal cycle.

$$cyclek = 1{:}\,M_{ec}, reproductioncyclel = 1{:}\,M_{qf}, andchemotaxiscyclei = 1{:}\,M_d$$

Step 4: Chemotaxis illness is conducted.

Step 5: The reproduction process occurs. Bacteria with low fitness values were killed, while those with high fitness values were duplicated extensively.

Step 6: The elimination-dispersal procedure is undertaken. Every bacterium

produces an O represents a random probability. This phase compares O to a pre-arranged movement possibility O_{fc}. If $O < O_{fc}$ the elimination-dispersal technique is done.

Step 7: Termination circumstances are tested. If the requirements are satisfied, the outcome is output. Otherwise, it goes back to step 4.

Step 8: Chemotaxis action involves two primary operators such as tumble and swim. Swimming means a constant motion towards optimal fitness. Equation (14) indicates the bacterial adaption value (I_{dd}). The following is a description of the equation: The product of $d(j)$ and $\Delta(j)$ is added to determine the updated value of $\theta^j(i+1,l,k)$, standardized by the square of the root of the average of $\Delta^S(j)$ and $\Delta(j)$. This update reflects an iterative modification to the variable θ^j in the subsequent iteration $(i+1)$, depending on the scaling variables involved and the present values of $\theta^j\ (i,\ l,\ k)$.

$$I_{dd}\big(\theta, O(i,l,k)\big) = \sum_{j=1}^{T} I_{dd}^{j}\left(\theta, \theta^j(i,l,k)\right)$$
$$= \sum_{j=1}^{T}\left[-c_{attract} exp\left(-\omega_{attract}\sum_{n=1}^{C}\left(\theta_n - \theta_n^j\right)^2\right)\right] \tag{14}$$
$$+ \sum_{j=1}^{T}\left[g_{repellant} exp\left(-\omega_{repellant}\sum_{n=1}^{C}\left(\theta_n - \theta_n^j\right)^2\right)\right]$$

where $d(j)$ indicates the starting paddling length in the selected direction and applies to any orientation. The ideas of attraction and repulsion are measured by factors that characterize their relative effects: the degree of attraction reveals the way a target is drawn in, while the breadth of attraction defines the range of influence. In contrast, the height and width of attraction determine the extent to which the repellent force operates. Together, these factors increase the method of optimization by allowing for subtler exploration in the answer space, resulting in greater efficiency in the resulting assistance for the decision model.

The Dynamic Bacterial Foraging Fine-Tuned Efficient Adaptive Boosting (DBF-EAdaBoost) hybrid method improves the performance of weak learners in EAdaBoost by constantly modifying sample weights and learner variables using a fitness function obtained from the Dynamic Bacterial Foraging (DBF) optimization technique. The major goal is to improve categorization performance in decision-making models used for health risk prediction. The technique makes use of data from 500 employees, including DNA sequences, biometrics, occupational health records, and demographic information. The DBF-EAdaBoost method uses swarm intelligence principles with adaptive boosting to achieve optimal or near-optimal solutions, increasing the model's ability to properly classify health risks while enhancing decision-making processes in workplace wellness situations. Algorithm 1 shows the hybrid model.

Algorithm 1 DBF-EAdaboost

Step 1: Initialize Sample Weights: Allocate a comparable weight to all observations in the dataset using Equation (16).

$$C_1(j) = \frac{1}{N} \quad (15)$$

Where N is the entire amount of samples.

Step 2: Train Weak Learner and Calculate Error Rate: Using incorrectly classified examples, train an ineffective classifier & calculate the mistake rate in Equation (17)

$$\epsilon_s = \frac{N_{err}}{N} \quad (16)$$

Where N_{err} is the quantity of misclassified models.

Step 3: Modify the Weak Learner's Value: Change the amount of weight of the weak learner depending on its error rate in Equation (18)

$$\alpha_s = \frac{1}{2} In\left(\frac{1-\varepsilon_s}{\varepsilon_s}\right) \quad (17)$$

Step 4: This process is used to update the position of bacteria. To modify the position, execute optimization with the bacterial foraging algorithm.

$$\theta^i(j+1) = \theta^i(j) + c(i)\frac{\Delta(i)}{\sqrt{\Delta(i)\Delta(i)}} \quad (18)$$

Where c(i) is the step size and Δ(i) is the direction vector.

Step 5: Merge Weak Learners for Ultimate Output: Combine the weak learners to produce the final classification result in Equation (19)

$$H(X) = \sigma(\sum_{i=1}^{t} \alpha_i h_i(X)) \quad (19)$$

Where $\sigma(.)$ is the sigmoid function, and $h_i(X)$ are the weak classifiers.

This structure shows the steps in the DBF-EAdaBoost example, with calculations for each step in the process. This DBF-EAdaBoost algorithm improves the accuracy of health risk prediction classification by combining optimal mouse foraging and Efficient Adaptive Boosting (EAdaBoost) methods. It uses bio-inspired optimization to constantly change sample weights and refine model parameters, focusing on complex issues. This hybrid approach enhances decision support models by combining vulnerable learners with improving virus coverage, resulting in accurate prediction of occupational health and bioinformatics ultimately improving the prediction of clinical outcomes.

4. Decision support model

The decision-support approach of this study focuses on the integration of bioinformatics with occupational health data to improve workplace health risk prediction. It uses a broad set of data including health markers, including employee contacts, for everyone in the company. Using this diverse set of data, the model hopes to identify future health problems posed by a workplace environment, providing a proactive approach to employee well-being. The model is constructed on the DBF-EAdaBoost approach, which dramatically increases classification performance. This innovative strategy combines dynamic bacterial foraging optimization with adaptive enhancement methods to enable the model to reliably fit health risk prediction based on individual genetic exposure at work. Furthermore, it speeds up the health risk assessment process, but it also promotes the sector to make data-driven decisions. Integrating multiple data sources helps provide a comprehensive picture of employee health, enhancing workplace safety and well-being. Implementation of the model could lead to more effective healthcare products and ultimately a healthier workforce. Furthermore, it highlights the value of a strong company culture that prioritizes employee well-being and paves the way for better

health outcomes and increased productivity in a rapidly changing workplace environment.

5. Experimental result

The experimental setup for the work includes Windows 11, Python 3.10, and PyTorch 2. x, which provide a stable environment for model development. The Ryzen 7 5800 X CPU is used for demanding calculations, while the Radeon RX 7900 XTX GPU speeds up data preparation, and feature extraction using Linear Discriminant Analysis (LDA), and model training. This high-performance architecture enables to effectively handle and evaluate bioinformatics and occupational health datasets from 500 employees, considerably improving the accuracy of health hazard categorization and facilitating effective workplace safety decisions.

Python 3.10 was utilized in the study to carry out the suggested methodology, and Python 3.10 is compatible with Pytorch 2.0 and incorporates the Windows 10 system setup. The experimental results reveal that classic models such as Random Forests or else, SVM, Decision Trees [21], and XGBoost [22,23] provide reasonable accuracy but face limitations such as poor generalization, noise sensitivity, and difficult parameter tweaking. Logistic Regression [23] issues linear assumptions in complicated datasets, and while CNN [23] is effective for image tasks, they are highly computational and require huge datasets.

ROC: The ROC (Receiver Operating Characteristic) curve is used to assess the efficiency of a classification algorithm. It compares the true positive rate (sensitivity) to the false positive rate (1-specificity) at various threshold levels, showing the trade-off between accurately recognizing ideal conditions and limiting false positives. This curve is critical in determining the model's capacity to differentiate between health hazards. By incorporating this review into the goal of creating a decision-support model utilizing bioinformatics along with occupational health data, both the precision and the accuracy of health risk forecasts can be improved, thus improving decision-making processes are seen in **Figure 2**.

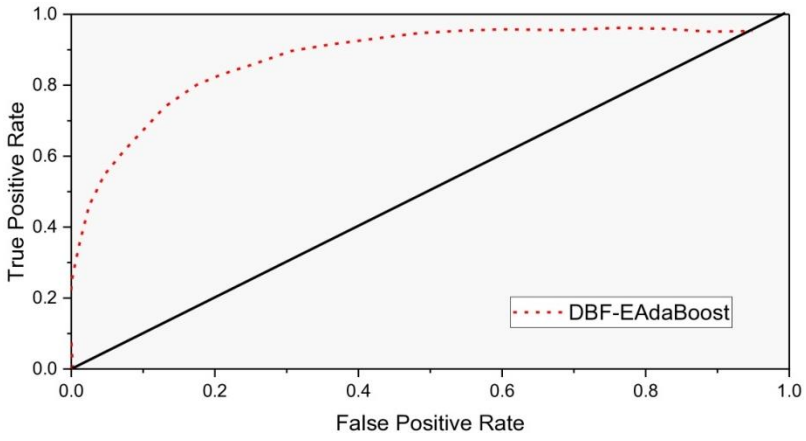

Figure 2. ROC-AUC Curve for DBF-EAdaBoost proposed method.

Accuracy and loss curve: The accuracy metric measures the rate of true predictions made by a model throughout training, which typically increases as the

model evolves. A rising accuracy graph demonstrates effective learning, but the loss meter measures error in sampling, with lower values indicating better model performance. Analyzing these graphs gives useful information on the model's advancement in academia, enabling performance improvements and assuring effective generalization of novel data. This method is critical to the goal of developing a strong decision-support framework that uses informatics and workplace health data to conduct reliable illness testing are displayed in **Figure 3a,b**.

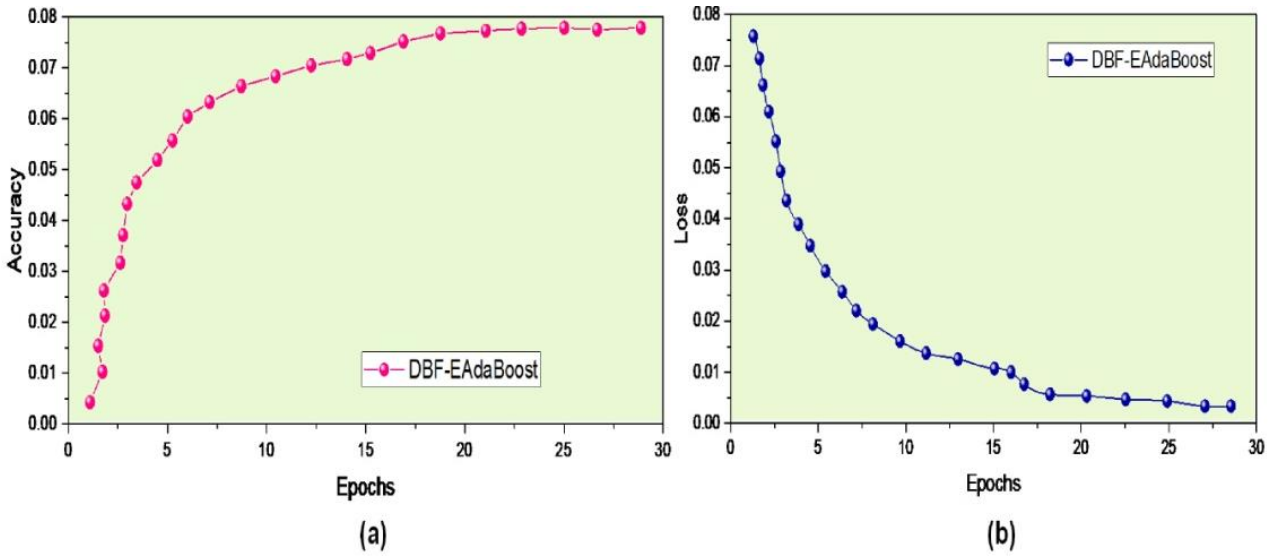

Figure 3. Proposed method DBF-EAdaBoost established. **(a)** accuracy curve; **(b)** Loss curve.

Figure 3a,**b** demonstrate the accuracy and loss of the DBE-EAdaBoost method during training. The accuracy curve (3a) illustrates the model's improving capacity to produce real prediction, whereas the loss curve (3b) shows a decrease in forecasting error. These curves operate together to analyze the model training development, ensuring that it is effectively generalized to new data and performs consistently for illness testing in occupational health contexts.

Accuracy: **Table 2** and **Figure 4** compare the accuracy of different models in classification tasks. Random Forests (0.890) and Decision Trees (0.880) excel at ensemble learning, whereas Support Vector Machine (0.884) achieves high classification accuracy using hyperplanes. XGBoost (0.866), a powerful boosting algorithm, has a slightly lesser accuracy, whilst Logistic Regression was (0.844) after, indicating its simpler linear approach. CNN (0.868) applies machine learning to difficult data but does not outperform classic algorithm design. The proposed DBF-EAdaBoost model achieves the best accuracy (0.93), demonstrating its improved capacity to refine predictions over traditional techniques.

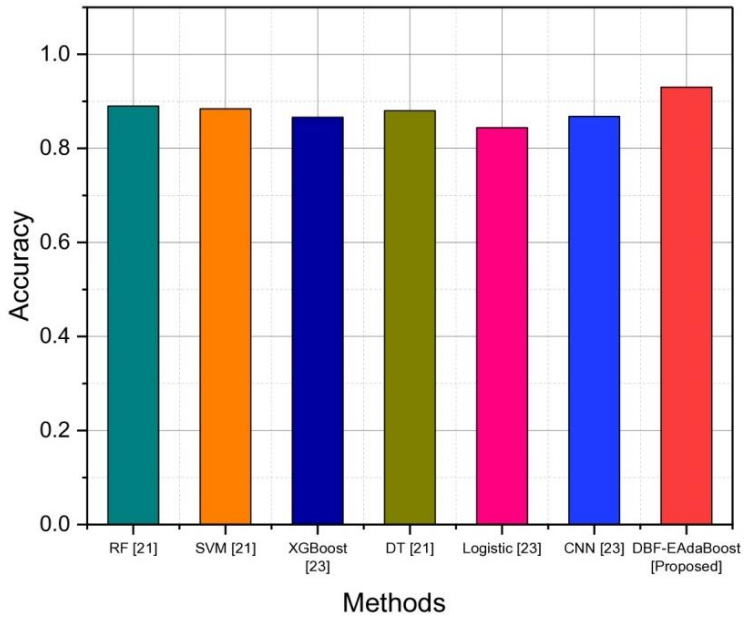

Figure 4. Comparison accuracy values of various predictive models vs. proposed method.

Table 2. Model performance comparison based on accuracy.

Models	Accuracy
RF [21]	0.890
SVM [21]	0.884
XGBoost [23]	0.866
DT [21]	0.880
Logistic [23]	0.844
CNN [23]	0.868
DBF-EAdaBoost [Proposed]	0.93

Precision: **Table 3** compares the precision of several models in forecasting health risks using datasets. **Figure 5** compares the levels of precision among various predictive models while highlighting the suggested DBF-EAdaBoost technique. Precision indicates that a model forecasts positive health results (true positives). Precision, which evaluates the accuracy of positive predictions, is high for models such as Support Vector Machine (0.984) and Decision Tree (0.980), implying that they accurately predict true positives. Random Forests (0.978) performs as well, however, XGBoost (0.633) has lesser precision, most likely because of model tuning or dataset features. The proposed model DBF-EAdaBoost has the highest precision (0.987), suggesting its greater capacity to consistently identify problems with health, minimize false positives, and improve prediction accuracy in health-related datasets.

Table 3. Model performance comparison based on precision.

Models	Precision
RF [21]	0.978
SVM [21]	0.984
XGBoost [22]	0.633
DT [21]	0.980
DBF-EAdaBoost [Proposed]	0.987

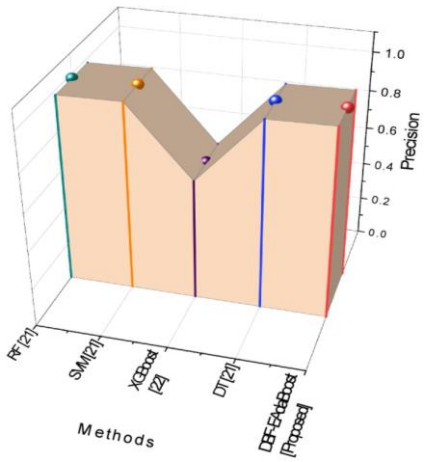

Figure 5. Comparison of precision values of various predictive models vs. proposed method.

Brier score: The Brier scores for several models assess the accuracy of their statistical predictions; lower numbers indicate greater performance. Both Random Forests were 0.107, and XGBoost has scores of 0.144, indicating that the likelihood of their predictions is similarly accurate. Logistic regression shows marginally increased performance, with a score of 0.121. In comparison, the Convolutional Neural Network (CNN) has the score (0.194), indicating less accurate probability forecasts. Particularly, the proposed model, DBF-EAdaBoost, has the best Brier score of 0.100, demonstrating its superior capacity to generate correct estimations of probability for health risk assessments, which is seen in **Table 4** and **Figure 6** compares the Brier scores for different models with suggested DBF-EAdaBoost approach. A lower Brier score indicates improved model effectiveness in probability determination.

Table 4. Comparison analysis of Brier score for various predictive models vs. proposed method.

Models	Brier score
RF [22]	0.107
XGBoost [22]	0.144
Logistic [23]	0.121
CNN [23]	0.194
DBF-EAdaBoost [Proposed]	0.100

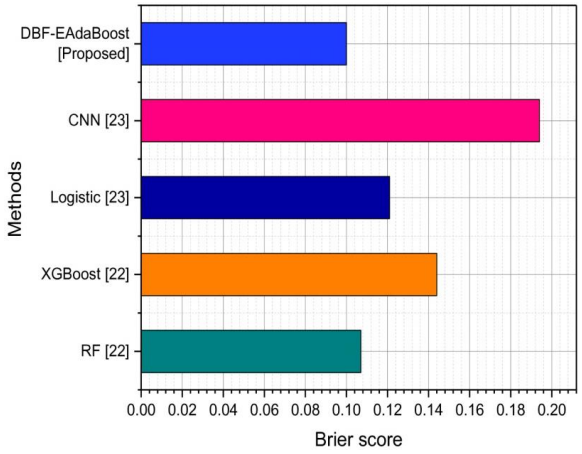

Figure 6. Classification of Brier scores for predictive models vs. proposed method.

AUC: The Area Under the Curve (AUC) ratings of various models indicate their ability to differentiate between positive and negative categories in binary classification problems. A higher AUC indicates improved model performance. The Support Vector Machine dominates with an AUC of 0.91, followed by the proposed model, DBF-EAdaBoost, which has an AUC of 0.92, proving that it's successful in classifying health concerns. Random Forests and Decision Trees both get a score of 0.90, showing high predictive skills. Whereas XGBoost (0.789), Logistic Regression (0.734), and CNN (0.724) have lower AUC values, indicating less effective classification performance, as shown in **Table 5** and **Figure 7** demonstrates the Area Under the Curve (AUC) evaluations for various methods and proposed technique, which assess their ability to correctly identify both positive and negative classifications. A greater AUC suggests improved performance.

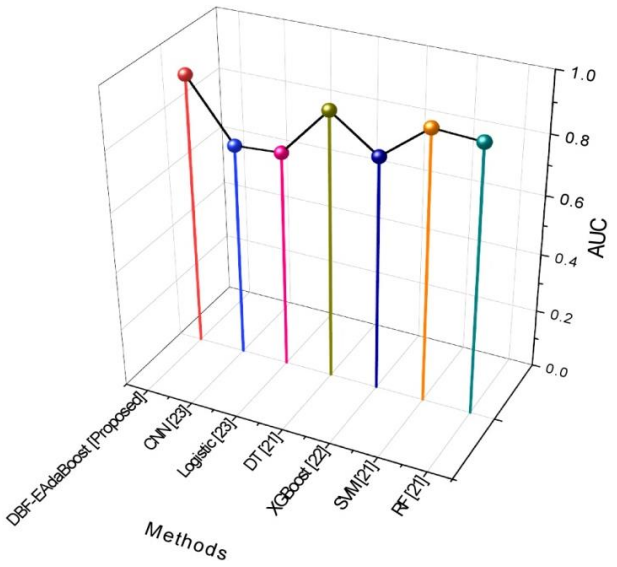

Figure 7. Classification of AUC for predictive models vs. proposed method.

Table 5. Performance finding of AUC for the proposed method.

Models	AUC
RF [21]	0.90
SVM [21]	0.91
XGBoost [22]	0.789
DT [21]	0.90
Logistic [23]	0.734
CNN [23]	0.724
DBF-EAdaBoost [Proposed]	0.92

Log loss: Log loss, also known as logistic loss, measures a classification model's performance in terms of expected probabilities. Reduced log loss values indicate improved model performance. Random Forests and XGBoost both have a log loss of 0.448, demonstrating comparable predictive accuracy when calculating probabilities. Convolutional Neural Networks (CNN) have a somewhat larger log loss (0.442), indicating less accurate predictions. Logistic regression is superior with a log loss of 0.398, while the recommended model, DBF-EAdaBoost, surpasses all others with a log loss of 0.314. This illustrates its superior capacity to provide a reliable estimate of the probability for safety evaluations, which is displayed in **Table 6** and **Figure 8** shows the Log Loss values of the traditional model with the suggested DBF-EAdaBoost method, where lower scores suggest greater performance in forecasting probabilities.

Table 6. Performance comparison result of log loss for the proposed method.

Models	Log loss
RF [22]	0.340
XGBoost [22]	0.448
Logistic [23]	0.398
CNN [23]	0.442
DBF-EAdaBoost [Proposed]	0.314

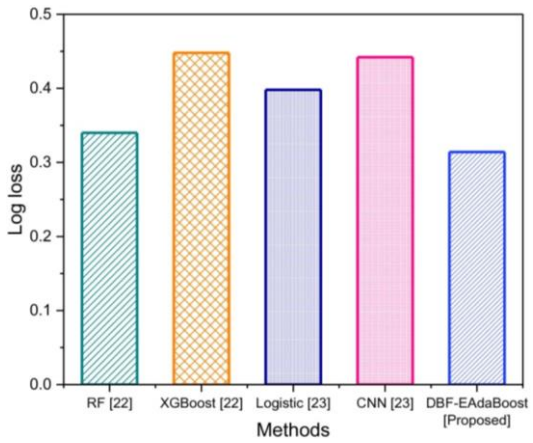

Figure 8. Comparison Log loss score values of various predictive methods vs. proposed method.

6. Discussion

It encompasses the integration of bioinformatics and occupational health data using advanced analytical techniques that can optimize the management of workplace health. Conventional techniques like RF, SVM, DT, XGBoost, logistic, and CNN have certain drawbacks when used in the context of a planning decision support model utilizing bioinformatics and occupational health data. RF [21,22] model was often referred to as a "black-box" model, which makes it challenging to completely understand how decisions are made, particularly in a high-risk industry such as occupational health. The SVM [21] is sensitive to large datasets since the entire dataset has to be stored in and operated upon to produce the desired hyperplane—which often means high memory and computation requirements, especially on bioinformatic datasets with many features. DT [21] was sensitive to overfitting, particularly if the tree is deep or if the dataset has noisy features. Various pruning techniques are employed with the risk of also limiting the model's ability to capture complex patterns. While XGBoost [22,23] performed well in many such cases, it was still prone to overfitting problems, especially if the model is highly complex and noisy data are used. Careful tuning of hyperparameters would be required; for example, the learning rate, tree depth, and subsampling of the general algorithm to reach the optimal performance of XGBoost, and this could indeed be quite expensive in terms of computation time. Logistic [23] assumed linear relationships between the features and the outcome. When the true relationship between features and outcome is non-linear, this could be limited, especially in datasets as complex as bioinformatics and health data. Training CNN [23] required a large number of labeled examples. Data might be scarce and difficult to label, limited the utilization of CNNs in bioinformatics and occupational health. The problems overcome in this study introduce the DBF-EAdaBoost algorithm proposed here to improve classification accuracy due to the effective integration of dynamic bacterial foraging optimization with adaptive boosting toward the enhancement of performance within both bioinformatics and occupational health data analysis. It has shown high precision and reliability, an approach that is qualified to produce forecasts regarding potential health problems based on both exposures at the workplace and genetic factors. It provides a more personalized and data-driven decision support model that further facilitates better-targeted interventions as well as better health outcomes among employees.

7. Conclusion

The planning decision support model effectively translates informatics and environmental data to address health issues in the contemporary workplace. Employing the enhanced version of the DBF technique, namely, the DBF-EAdaBoost, the model enhances the accuracy and precision of disease risk estimates essential to direct treatments like glucose level, blood pressure, etc. These novel approaches allow for a greater awareness of the way specific genetic features affect or are influenced by employment situations, resulting in more effective health interventions. The proposed DBF-EAdaBoost achieves the significant outcomes of accuracy (0.93), AUC (0.92), log loss (0.314), Brier score (0.100), and precision

(0.987). The research involves more than one performance measure and assesses the model against standard techniques in predicting potential health challenges, critical for health management. This capability facilitates the decision-making of public health officials and occupational health specialists for staff health and workplace safety improvement. Moreover, the integrated use of bioinformatics allows for the constant updating of information in the model to account for changes in conditions at work or new pathological threats. From there, that approach can be tailored and applied in all kinds of industries, making it a very versatile model for health risk assessment and management. Lastly, it helps to have a healthy workforce because it assists regarding the current standards for the kind of wellness programs needed in organizations.

Limitation and future scope

The main limitation of the developed model is its dependency on acquiring and utilizing high-quality comprehensive bioinformatics and occupational health data. Such data are not often available or standardized, which limits the use of such models. Future work will include incorporating real-time data streams, further improvement in adaptability, and integration with broader genetic, environmental, and lifestyle factors to further improve the precision and applicability of work-related health interventions.

Author contributions: Writing—original draft preparation, JL; writing—review and editing, WL and FC. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Kuipers, S.J., Nieboer, A.P. and Cramm, J.M., 2020. Views of patients with multi-morbidity on what is important for patient-centered care in the primary care setting. BMC Family Practice, 21, pp.1-12.https://doi.org/10.1186/s12875-020-01144-7
2. Kwame, A. and Petrucka, P.M., 2021. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nursing, 20(1), p.158.https://doi.org/10.1186/s12912-021-00684-2
3. Keij, S.M., van Duijn-Bakker, N., Stiggelbout, A.M. and Pieterse, A.H., 2021. What makes a patient ready for shared decision making? A qualitative study. Patient Education and Counseling, 104(3), pp.571-577.https://doi.org/10.1016/j.pec.2020.08.031
4. Al-Jaroodi, J., Mohamed, N. and Abukhousa, E., 2020. Health 4.0: on the way to realizing the healthcare of the future. Ieee Access, 8, pp.211189-211210.https://doi.org/10.1109/ACCESS.2020.3038858
5. Zeadally, S. and Bello, O., 2021. Harnessing the power of Internet of Things based connectivity to improve healthcare. Internet of Things, 14, p.100074.https://doi.org/10.1016/j.iot.2019.100074
6. Sharma, N., Dev, J., Mangla, M., Wadhwa, V.M., Mohanty, S.N. and Kakkar, D., 2021. A heterogeneous ensemble forecasting model for disease prediction. New Generation Computing, pp.1-15.https://doi.org/10.1007/s00354-020-00119-7
7. Javaid, M., Haleem, A. and Singh, R.P., 2024. Health informatics to enhance the healthcare industry's culture: An extensive analysis of its features, contributions, applications and limitations. Informatics and Health.https://doi.org/10.1016/j.infoh.2024.05.001
8. Marwaha, S., Knowles, J.W. and Ashley, E.A., 2022. A guide for the diagnosis of rare and undiagnosed disease: beyond the

exome. Genome medicine, 14(1), p.23.https://doi.org/10.1186/s13073-022-01026-w

9. Panayides, A.S., Amini, A., Filipovic, N.D., Sharma, A., Tsaftaris, S.A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T. and Huang, K., 2020. AI in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics, 24(7), pp.1837-1857.https://doi.org/10.1109/JBHI.2020.2991043
10. Soh, E., Tsai, J.H.C., Boutain, D.M. and Pike, K., An intersectional analysis of the health status, work conditions, and nonwork conditions of the US working‐classed across class, sex, race, and nativity identities. American Journal of Industrial Medicine.https://doi.org/10.1002/ajim.23663
11. Das, S., Khanwelkar, D.R. and Maiti, J., 2024. A semi-automated coding scheme for occupational injury data: An approach using Bayesian decision support system. Expert Systems with Applications, 237, p.121610.https://doi.org/10.1016/j.eswa.2023.121610
12. Casal-Guisande, M., Comesaña-Campos, A., Dutra, I., Cerqueiro-Pequeño, J. and Bouza-Rodríguez, J.B., 2022. Design and development of an intelligent clinical decision support system applied to the evaluation of breast cancer risk. Journal of personalized medicine, 12(2), p.169.https://doi.org/10.1109/TMSCS.2017.2710194
13. Tutun, S., Johnson, M.E., Ahmed, A., Albizri, A., Irgil, S., Yesilkaya, I., Ucar, E.N., Sengun, T. and Harfouche, A., 2023. An AI-based decision support system for predicting mental health disorders. Information Systems Frontiers, 25(3), pp.1261-1276.https://doi.org/10.1007/s10796-022-10282-5
14. Reska, D., Czajkowski, M., Jurczuk, K., Boldak, C., Kwedlo, W., Bauer, W., Koszelew, J. and Kretowski, M., 2021. Integration of solutions and services for multi-omics data analysis towards personalized medicine. biocybernetics and biomedical engineering, 41(4), pp.1646-1663.https://doi.org/10.1016/j.bbe.2021.10.005
15. Patel, V., Chesmore, A., Legner, C.M. and Pandey, S., 2022. Trends in workplace wearable technologies and connected‐worker solutions for next‐generation occupational safety, health, and productivity. Advanced Intelligent Systems, 4(1), p.2100099.https://doi.org/10.1002/aisy.202100099
16. Viet, S.M., Falman, J.C., Merrill, L.S., Faustman, E.M., Savitz, D.A., Mervish, N., Barr, D.B., Peterson, L.A., Wright, R., Balshaw, D. and O'Brien, B., 2021. Human Health Exposure Analysis Resource (HHEAR): A model for incorporating the exposome into health studies. International journal of hygiene and environmental health, 235, p.113768.https://doi.org/10.1016/j.ijheh.2021.113768
17. Saravi, B., Hassel, F., Ülkümen, S., Zink, A., Shavlokhova, V., Couillard-Despres, S., Boeker, M., Obid, P. and Lang, G.M., 2022. Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. Journal of Personalized Medicine, 12(4), p.509.https://doi.org/10.3390/jpm12040509
18. Kohn, M.S., Topaloglu, U., Kirkendall, E.S., Dharod, A., Wells, B.J. and Gurcan, M., 2022. Creating learning health systems and the emerging role of biomedical informatics. Learning Health Systems, 6(1), p.e10259.https://doi.org/10.1002/lrh2.10259
19. Gallagher, D., Zhao, C., Brucker, A., Massengill, J., Kramer, P., Poon, E.G. and Goldstein, B.A., 2020. Implementation and continuous monitoring of an electronic health record embedded readmissions clinical decision support tool. Journal of personalized medicine, 10(3), p.103.https://doi.org/10.3390/jpm10030103
20. Şık, A.S., Aydınoğlu, A.U. and Son, Y.A., 2021. Assessing the readiness of Turkish health information systems for integrating genetic/genomic patient data: System architecture and available terminologies, legislative, and protection of personal data. Health Policy, 125(2), pp.203-212.https://doi.org/10.1016/j.healthpol.2020.12.004
21. Khairuddin, M.Z.F., Lu Hui, P., Hasikin, K., Abd Razak, N.A., Lai, K.W., Mohd Saudi, A.S. and Ibrahim, S.S., 2022. Occupational injury risk mitigation: machine learning approach and feature optimization for smart workplace surveillance. International journal of environmental research and public health, 19(21), p.13962.https://doi.org/10.3390/ijerph192113962
22. Zhao, Z., Lu, H., Meng, R., Si, Z., Wang, H., Wang, X., Chen, J., Zheng, Y., Wang, H., Hu, J. and Zhao, Z., 2024. Risk factor analysis and risk prediction study of obesity in steelworkers: model development based on an occupational health examination cohort dataset. Lipids in Health and Disease, 23(1), p.10. https://doi.org/10.1186/s12944-023-01994-x
23. Zheng, Z., Si, Z., Wang, X., Meng, R., Wang, H., Zhao, Z., Lu, H., Wang, H., Zheng, Y., Hu, J. and He, R., 2023. Risk prediction for the development of hyperuricemia: Model development using an occupational health examination dataset. International Journal of Environmental Research and Public Health, 20(4), p.3411. https://doi.org/10.3390/ijerph20043411