

Гигиена / Hygiene

Научная статья

Original article

DOI: [10.12731/2658-6649-2026-18-3-1815](https://doi.org/10.12731/2658-6649-2026-18-3-1815)EDN: [XITOHH](https://www.edn.ru/XITOHH)

UDC 613.6:614.2



Neural network classification of a comprehensive occupational health risk index for agricultural workers in the peripheral zones of a megacity: Data from the Leningrad Region

Li Yin¹, O.T. Ergunova², A.B. Titov^{2,3}, A.G. Somov²¹Harbin Institute of Technology, Harbin, China²Peter the Great St. Petersburg Polytechnic University,
Saint Petersburg, Russian Federation³Financial University under the Government of the Russian Federation
(St. Petersburg branch), Saint Petersburg, Russian Federation

Abstract

Background. Agricultural workers in periurban zones of large metropolitan areas are exposed to a complex combination of occupational, ecological and socio-occupational stressors that substantially elevate the burden of work-related diseases. Preventive medicine frameworks require quantitative, data-driven instruments capable of stratifying individual health risk under conditions of high inter-individual variability and incomplete administrative records. However, existing categorical risk scoring systems treat exposure domains independently and rely on manual expert evaluation, limiting their scalability.

Purpose. To develop and validate a feedforward artificial neural network (ANN) model for automated classification of an expert-constructed composite occupational health risk index (OHRI) among agro-industrial workers in the Leningrad Region, and to assess the relative contribution of ecological, physiological and socio-occupational predictor domains to risk classification through permutation importance analysis.

Methods. A cross-sectional dataset comprising 1,248 agricultural workers from 14 enterprises of the Leningrad Region (index examinations conducted between 2021 and 2025) was assembled from periodic medical examinations, workplace sanitary inspections, and regional environmental monitoring records. The OHRI was constructed as a weighted composite of three domain scores – Ecological Exposure Score (EES), Physiological Strain Score (PSS), and Social-Labor Stress Score (SLSS) – with weights derived through Analytic Hierarchy Process (AHP) expert elicitation. A feedforward ANN with two hidden layers (architecture: 22–64–32–1) was implemented in PyTorch and trained using a 70/15/15 stratified split. The ANN was benchmarked against logistic regression with elastic net regularization (LR-EN) and a gradient boosting classifier (GBC, XG-

Boost). Model performance was evaluated by AUC-ROC, sensitivity, specificity, F1 score, and Brier score on the holdout test set. Enterprise-aware group splitting was applied to mitigate data leakage arising from the clustered data structure.

Results. The ANN model achieved an AUC-ROC of 0.872 (95 % CI: 0.831–0.913), outperforming logistic regression (AUC 0.803, DeLong $p < 0.001$) and performing comparably to gradient boosting (AUC 0.858, DeLong $p = 0.214$). The five most important predictors by permutation importance were: cumulative pesticide exposure index (mean importance = 0.083), heart rate variability SDNN (0.071), ambient PM_{2.5} concentration at the worksite (0.063), subjective labor intensity score (0.054), and years of employment in the sector (0.049). On the holdout set, the ANN achieved sensitivity of 0.806, specificity of 0.862, and an overall accuracy of 0.843.

Conclusion. A feedforward neural network can classify an expert-constructed composite occupational health risk index with good discriminatory accuracy and may serve as a scalable digital tool for automated risk stratification in agro-industrial enterprises. The ecological domain contributed the largest share of predictor importance, followed by physiological and socio-occupational domains. Prospective validation against incident occupational disease outcomes is required before the model can be recommended for clinical deployment.

Keywords: artificial neural network; occupational health risk index; agricultural workers; preventive medicine; periurban area; ecological monitoring; socio-occupational factors; analytic hierarchy process; composite risk index; machine learning; Leningrad Region

For citation. Li, Yin, Ergunova, O. T., Titov, A. B., & Somov, A. G. (2026). Neural network classification of a comprehensive occupational health risk index for agricultural workers in the peripheral zones of a megacity: Data from the Leningrad Region. *Siberian Journal of Life Sciences and Agriculture*, 18(3), 323-348. <https://doi.org/10.12731/2658-6649-2026-18-3-1815>

Нейросетевая классификация комплексного индекса профессионального риска здоровью работников агропромышленного комплекса периферийных зон мегаполиса: данные Ленинградской области

Ли Инь¹, О.Т. Ергунова², А.Б. Титов^{2,3}, А.Г. Сомов²

¹Харбинский технологический институт, Харбин, Китай

²Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

³Финансовый университет при Правительстве РФ (Санкт-Петербургский филиал), Санкт-Петербург, Российская Федерация

Аннотация

Обоснование. Работники аграрного сектора периферийных зон мегаполисов подвергаются комплексному воздействию профессиональных, экологических и

социально-трудовых стрессоров, что существенно увеличивает бремя профессиональной заболеваемости. Существующие категориальные системы оценки профессионального риска рассматривают домены экспозиции независимо и основаны на ручной экспертной оценке, что ограничивает их масштабируемость.

Цель. Разработка и валидация модели на основе многослойного персептрона (искусственной нейронной сети, ИНС) для автоматизированной классификации экспертно-конструированного комплексного индекса профессионального риска здоровью (ИПРЗ) работников агропромышленных предприятий Ленинградской области, а также оценка относительного вклада экологических, физиологических и социально-трудовых предикторов в классификацию риска методом пермутационной важности.

Методы. Сформирована поперечная выборка из 1 248 работников АПК 14 предприятий Ленинградской области (индексные осмотры за 2021–2025 гг.) на основании данных периодических медицинских осмотров (ПМО), санитарно-гигиенических характеристик рабочих мест и регионального экологического мониторинга. ИПРЗ рассчитан как взвешенная композиция трёх доменных оценок (экологической, физиологической, социально-трудовой) с весами, полученными методом анализа иерархий (АНР). ИНС с двумя скрытыми слоями (архитектура: 22–64–32–1) реализована в PyTorch, обучена в разбивке 70/15/15 с учётом кластерной структуры по предприятиям и сопоставлена с логистической регрессией и градиентным бустингом (XGBoost).

Результаты. Модель ИНС достигла $AUC-ROC = 0,872$ (95 % ДИ: 0,831–0,913), превзойдя логистическую регрессию (AUC 0,803; DeLong $p < 0,001$) и показав сопоставимую с градиентным бустингом точность (AUC 0,858; DeLong $p = 0,214$). Пять наиболее значимых предикторов (по пермутационной важности): кумулятивный индекс воздействия пестицидов (0,083), вариабельность сердечного ритма SDNN (0,071), концентрация $PM_{2,5}$ на рабочем месте (0,063), субъективная оценка трудовой интенсивности (0,054), стаж работы в отрасли (0,049). Чувствительность модели – 0,806, специфичность – 0,862, общая точность – 0,843.

Заключение. ИНС обеспечивает автоматизированную классификацию экспертно-конструированного комплексного индекса профессионального риска с хорошей дискриминационной точностью. Для клинического внедрения модели необходима проспективная валидация с использованием подтверждённых клинических исходов профессиональных заболеваний. Важно подчеркнуть, что текущая модель классифицирует экспертно-конструированный индекс, а не независимый клинический исход, что ограничивает клиническую интерпретируемость результатов.

Ключевые слова: искусственная нейронная сеть; индекс профессионального риска здоровью; работники АПК; профилактическая медицина; периферия мега-

полиса; экологический мониторинг; социально-трудовые факторы; метод анализа иерархий; композитный индекс риска; машинное обучение; Ленинградская область

Для цитирования. Ли, Инь, Ергунова, О. Т., Титов, А. Б., & Сомов, А. Г. (2026). Нейросетевая классификация комплексного индекса профессионального риска здоровью работников агропромышленного комплекса периферийных зон мегаполиса: данные Ленинградской области. *Siberian Journal of Life Sciences and Agriculture*, 18(3), 323-348. <https://doi.org/10.12731/2658-6649-2026-18-3-1815>

1. Introduction

The agro-industrial sector of the Leningrad Region constitutes a structurally significant component of the labor market of the St. Petersburg metropolitan agglomeration. With over 47,000 persons employed in agricultural production and food processing enterprises within a 150 km radius of the city center [1], the region is characterized by a demographically aging workforce concentrated in periurban rural territories, subject simultaneously to the ecological pressures of intensive farming and the socio-occupational tensions generated by proximity to a megacity economy.

Occupational health outcomes in this population are shaped by an intersection of exposures that are rarely modeled jointly: chronic pesticide contamination of soil and air [2], physical ergonomic load associated with crop cultivation and livestock operations [3], cardiovascular strain induced by irregular shift patterns [4], and the psychological burden arising from perceived job insecurity and wage differentials relative to the urban core [5]. Epidemiological surveillance data for the Leningrad Region indicate that the standardized incidence of work-related musculoskeletal disorders, respiratory diseases, and cardiovascular pathologies among agricultural workers exceeds the regional population average by factors of 1.8, 2.1, and 1.6, respectively [6], a pattern corroborated by recent studies of combined occupational exposures in Russian agricultural settings [28].

Preventive medicine frameworks have traditionally responded to this complexity through the application of categorical risk scoring systems (e.g., the Russian occupational risk classes R 2.2.2006-05) or single-exposure dose–response modeling. Such approaches are limited by their inability to capture non-linear interactions between predictors and their poor performance under the data sparsity that characterizes workplace health records in small and medium agricultural enterprises [7, 25]. Machine learning methods, and in particular artificial neural networks (ANNs), offer a methodological alternative capable of modeling complex multivariate relationships without prior specification of

functional form [8]. In the context of Industry 4.0, the integration of data-driven approaches into occupational health management has been identified as a critical priority [9]. While these methods do not inherently resolve the challenges of exposure measurement or outcome definition, they may improve the scalability of existing expert-based risk classification frameworks by automating the integration of heterogeneous data sources.

It is important to note that the present study does not aim to discover novel biological risk pathways, but rather to construct a scalable digital instrument capable of replicating expert-derived composite risk assessments on the basis of routinely collected data. In this respect, the neural network approach serves as a tool for automating the expert consensus encoded in the OHRI construction, extending the practical reach of occupational health surveillance without requiring the manual involvement of specialist panels for each individual worker evaluation.

The specific objectives of the study are: (1) to construct a composite Occupational Health Risk Index (OHRI) integrating ecological, physiological and socio-occupational dimensions for agricultural workers; (2) to develop and validate an ANN model for automated classification of OHRI categories; (3) to assess the relative importance of predictor domains through permutation importance analysis; and (4) to benchmark ANN performance against conventional statistical classifiers.

2. Materials and Methods

2.1. Study Design, Population and Data Sources

A cross-sectional observational study was conducted using administrative and clinical data from 1,248 workers employed at 14 agro-industrial enterprises of the Leningrad Region. For each participant, the index observation was defined as the most recent complete periodic medical examination (PME) record within the period 2021–2025, with corresponding ecological and workplace exposure data matched to the calendar year of the examination. Thus, each worker contributed a single cross-sectional record; no repeated measurements were used. The study followed the STROBE guidelines for cross-sectional studies.

Enterprises were selected to represent the principal agricultural sub-sectors of the region: grain and feed crop production ($n = 4$ enterprises, 342 workers), greenhouse vegetable cultivation ($n = 3$, 298 workers), dairy and beef livestock ($n = 4$, 387 workers), and fish aquaculture ($n = 3$, 221 workers). Workers aged 18–65 years with a minimum of one year of continuous employment at the study enterprise were eligible for inclusion.

Three data streams were integrated: (i) individual health records from mandatory PME, conducted in accordance with Order of the Ministry of Health of the Russian Federation No. 29n of 28.01.2022; (ii) workplace sanitary and hygienic characterization protocols (SHCP) compiled by regional Rospotrebnadzor inspectorates; (iii) ambient environmental quality data from 22 automated monitoring stations of the Federal State Monitoring Network operated by the Leningrad Regional Environmental Committee (LenKomEco). Data linkage was performed at the enterprise level using a pseudonymized identifier: environmental monitoring values from the station nearest to each enterprise were assigned to all workers of that enterprise. This enterprise-level linkage introduces a clustered data structure that was accounted for in the validation strategy (see Section 2.5).

Sample flow and missing data. Of 1,647 workers initially screened, 399 (24.2 %) were excluded: 187 (11.4 %) due to incomplete PME records (missing spirometry, HRV or blood pressure data), 124 (7.5 %) due to less than one year of employment at the index enterprise, and 88 (5.3 %) due to absence of matched SHCP data. No imputation was performed; complete-case analysis was used. The excluded group did not differ significantly from the included group in age ($p = 0.34$), sex ($p = 0.61$), or enterprise distribution ($\chi^2 p = 0.28$), suggesting that selection bias due to exclusion was limited.

2.2. Composite Occupational Health Risk Index (OHRI)

The OHRI was calculated as a weighted composite of three domain scores: (a) Ecological Exposure Score (EES), (b) Physiological Strain Score (PSS), and (c) Social-Labor Stress Score (SLSS).

Domain score construction. Each domain score was computed as the arithmetic mean of its constituent indicator z-scores, with sign conventions set so that higher values indicate greater risk. Specifically:

- *EES* comprised six indicators: cumulative pesticide exposure index (PEI), soil heavy metal contamination index (HMCI), ambient $PM_{2.5}$, SO_2 , and NO_2 annual mean concentrations, and workplace noise level. All were positively oriented (higher = greater risk).
- *PSS* comprised seven indicators: resting heart rate (+), HRV SDNN (inverted: -), systolic blood pressure (+), diastolic blood pressure (+), BMI (+), FEV_1/FVC ratio (inverted: -), and hemoglobin concentration (inverted: -). The signs in parentheses indicate the direction of the z-score relative to risk.
- *SLSS* comprised nine indicators: years of employment (+), weekly working hours (+), shift irregularity index (+), subjective labor intensity score (+), occupational category (binary: manual = 1, supervisory = 0), monthly wage

relative to regional median (inverted: –), commuting time (+), job satisfaction score (inverted: –), and perceived employment stability score (inverted: –).

Each raw indicator was standardized to a z-score using the sample mean and standard deviation. Domain scores (mean of constituent z-scores) were then normalized to the [0, 1] interval via min–max scaling across the sample.

Domain weighting. Weights were assigned through expert elicitation using the Analytic Hierarchy Process (AHP) [20] with a panel of 11 occupational health physicians and 6 agricultural ecologists. Pairwise comparisons were collected individually via structured questionnaires and aggregated using the geometric mean method. The resulting weights were: $w_{EES} = 0.38$, $w_{PSS} = 0.35$, $w_{SLSS} = 0.27$. The overall consistency ratio (CR) was 0.04, below the conventional threshold of 0.10, indicating acceptable consistency of expert judgments. The composite formula is:

$$OHRI = 0.38 \cdot EES + 0.35 \cdot PSS + 0.27 \cdot SLSS$$

Workers with $OHRI \geq 0.60$ were classified as high-risk (binary outcome: $Y = 1$); those with $OHRI < 0.60$ as low-to-moderate risk ($Y = 0$). This threshold was selected on the basis of the distribution of OHRI values (natural break in the kernel density estimate) and cross-referenced with published occupational morbidity estimates for the region [6], yielding a high-risk prevalence of 29.4 % in the study sample. As a sensitivity analysis, classification performance was also evaluated at alternative thresholds (0.50, 0.55, 0.65, 0.70) to assess the robustness of findings to the threshold choice (see Supplementary Table S2).

Important methodological note. Because the OHRI is a composite index constructed from the same data domains that serve as model predictors, the classification task is not equivalent to prediction of an independent clinical outcome. Rather, the ANN model serves as an automated approximation of the expert-defined OHRI formula, enabling rapid individual risk stratification from raw data without manual computation. This design choice is discussed further in the Limitations section.

2.3. Predictor Variables

Twenty-two predictor variables were assembled from the three data streams.

Ecological domain (6 variables): cumulative pesticide exposure index (PEI, $\text{mg}/\text{m}^3 \cdot \text{years}$), soil heavy metal contamination index (HMCI), ambient $\text{PM}_{2.5}$ ($\mu\text{g}/\text{m}^3$), SO_2 ($\mu\text{g}/\text{m}^3$), and NO_2 ($\mu\text{g}/\text{m}^3$) annual mean concentrations at the enterprise location, and noise level (dB(A)).

- *PEI* was calculated for each worker as the product of the mean airborne pesticide concentration (in mg/m^3) measured at the enterprise during

SHCP inspections (averaged across the most recent three inspection rounds) and the worker's years of employment at that enterprise. The calculation included all organophosphate and pyrethroid compounds measured in workplace air samples.

- *HMCI* was calculated as the summation pollution index (Z_c) according to the methodology of Saet et al. (1990), based on soil sampling data from the enterprise territory reported in SHCP records. The Z_c index represents the sum of concentration coefficients for individual heavy metals (Pb, Cd, Zn, Cu, Ni) relative to regional background levels, minus the number of elements plus one.
- Ambient air quality variables ($PM_{2.5}$, SO_2 , NO_2) were obtained from the nearest automated monitoring station and represent calendar-year means for the year of the index PME. Because these values were assigned at the enterprise level, all workers of a given enterprise share the same environmental exposure values. This ecological linkage may underestimate within-enterprise exposure variability, particularly between indoor (e.g., greenhouse) and outdoor (e.g., field) work environments.

Physiological domain (7 variables): resting heart rate (bpm), heart rate variability SDNN (ms), systolic and diastolic blood pressure (mmHg), body mass index (BMI, kg/m^2), spirometric FEV_1/FVC ratio, and hemoglobin concentration (g/dL). All measurements were obtained during the mandatory PME.

Socio-occupational domain (9 variables): years of employment in the sector, weekly working hours, shift irregularity index (ratio of non-standard to standard shifts over the preceding 3 months), subjective labor intensity score (5-point Likert scale, self-administered questionnaire), occupational category (manual/supervisory), monthly gross wage relative to regional median (%), commuting time (minutes), job satisfaction score (5-point Likert), and perceived employment stability score (5-point Likert).

2.4. Neural Network Architecture and Training

A feedforward ANN with two hidden layers was implemented in PyTorch 2.1. Architecture selection followed a systematic grid search over hidden layer sizes $\{(16, (32, (64, (32, 16), (64, 32), (128, 64)\}$, learning rates $\{0.001, 0.005, 0.01\}$, L2 regularization coefficients $\{0.0001, 0.001, 0.01\}$, and dropout rates $\{0.0, 0.15, 0.25, 0.35\}$, evaluated by five-fold stratified cross-validation on the training set.

The final architecture comprised an input layer of 22 nodes, two hidden layers of 64 and 32 neurons respectively (ReLU activation), and a single sigmoid output node (architecture: 22–64–32–1). Dropout regularization (rate = 0.25) was applied after each hidden layer. The Adam optimizer was used with a learning rate of 0.001 and L2 weight decay $\lambda = 0.0001$. Training was terminated by early stopping with a patience of 20 epochs, monitoring the validation loss. The final model converged after 142 epochs (of a maximum 500), with training and validation loss curves showing stable convergence without evidence of overfitting (see Supplementary Figure S2).

Preprocessing. All continuous predictor variables were standardized to zero mean and unit variance using the training set statistics. The binary variable (occupational category) was left unscaled. Standardization parameters were computed on the training set only and applied to validation and test sets to prevent data leakage.

Class imbalance. Synthetic minority oversampling (SMOTE) [19] was applied exclusively within each fold of the cross-validation procedure during hyperparameter tuning, and on the full training set (after final hyperparameter selection) prior to final model fitting. SMOTE was never applied to validation or test data. To preserve the clustered data structure, SMOTE was applied within enterprises: synthetic observations were generated only by interpolating between workers of the same enterprise, thereby avoiding the creation of artificial records with inconsistent enterprise-level environmental exposure values.

Classification threshold. The optimal probability threshold for binary classification was determined by maximizing the Youden index ($J = \text{sensitivity} + \text{specificity} - 1$) on the validation set. The resulting threshold (0.47) was applied to the holdout test set for computation of all confusion-matrix-derived metrics.

2.5. Benchmark Models, Validation Strategy and Performance Metrics

Benchmark models. The ANN was benchmarked against two comparators: (1) a logistic regression model with elastic net regularization (LR-EN) tuned via five-fold cross-validation; and (2) a gradient boosting classifier (GBC, XGBoost 1.7) [10] with hyperparameters tuned via the same cross-validation procedure over the following grid: number of estimators {100, 200, 300}, maximum depth {3, 4, 5, 6}, learning rate {0.01, 0.05, 0.1}, and subsample ratio {0.7, 0.8, 1.0}. All models received the same standardized feature set.

Enterprise-aware validation. Given the clustered data structure (workers nested within 14 enterprises sharing identical environmental exposure values), a naive random train/test split would risk data leakage: the model could learn

to recognize enterprise-level exposure patterns rather than individual risk profiles. To mitigate this, the primary validation strategy used a group-aware split in which entire enterprises were assigned to training (10 enterprises, $n = 874$), validation (2 enterprises, $n = 187$), or test (2 enterprises, $n = 187$) sets, ensuring that no enterprise appeared in more than one partition. As a sensitivity analysis, a leave-one-enterprise-out cross-validation was also performed.

Performance metrics. Performance on the holdout test set was quantified by: area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, Brier score, and calibration slope and intercept. 95 % confidence intervals for AUC-ROC were computed using 2,000-replicate stratified bootstrap. Statistical significance of pairwise AUC differences was assessed using the two-sided DeLong test [18] ($\alpha = 0.05$). Predictor importance was estimated by permutation importance over 100 replications on the test set.

Calibration assessment. Calibration was evaluated using calibration plots (observed event rate vs. predicted probability in decile bins), the Hosmer–Lemeshow goodness-of-fit test, and calibration-in-the-large (calibration slope and intercept). Brier score decomposition into reliability, resolution and uncertainty components was performed.

2.6. Multicollinearity and Statistical Analysis

Multicollinearity. Variance inflation factors (VIF) were computed for all 22 predictors. Predictors with $VIF > 10$ were flagged. Pairwise Pearson correlations were examined for all predictor pairs; correlation matrices are presented in Supplementary Table S1. Systolic and diastolic blood pressure showed the highest pairwise correlation ($r = 0.72$); $PM_{2.5}$ and NO_2 were moderately correlated ($r = 0.58$). All VIF values remained below 5.0, and no predictors were excluded on the basis of collinearity.

Descriptive statistics are reported as means $\pm$ standard deviations for continuous variables and as counts (percentages) for categorical variables. Group differences between high- and low-risk workers were tested by independent-samples t-test (continuous) or chi-square test (categorical), with $p < 0.05$ considered statistically significant. It should be noted that because the high-risk and low-risk groups are defined by the OHRI, which is itself constructed from the predictor variables, significant group differences are partially tautological and do not constitute independent evidence of predictor-outcome associations.

Pearson correlation coefficients were computed between individual predictors and the continuous OHRI. All analyses were performed in Python 3.10 (Py-

Torch 2.1, Scikit-learn 1.3, Pandas 2.0, SciPy 1.11) and R 4.3 (pROC package for DeLong tests, CalibrationCurves package).

2.7. Ethical Approval

The study was approved by the Ethics Committee of Peter the Great St. Petersburg Polytechnic University (Protocol No. [INSERT BEFORE SUBMISSION], dated [INSERT DATE].2025). All data were obtained from administrative records (mandatory periodic medical examinations, sanitary inspection protocols, and environmental monitoring databases) and were fully pseudonymized prior to analysis. Individual informed consent was waived by the Ethics Committee given the retrospective, observational design and the use of de-identified administrative data. The study was conducted in accordance with the Declaration of Helsinki (2013 revision) and Russian Federal Law No. 152-FZ “On Personal Data.”

3. Results

3.1. Descriptive Characteristics of the Study Population

Table 1 presents the sociodemographic and occupational characteristics of the 1,248 study participants. The mean age was 42.7 ± 11.4 years. Male workers constituted 58.3 % of the sample. Livestock enterprise workers showed the highest proportion of high-risk classifications (38.1 %), followed by greenhouse cultivation (32.4 %), grain production (26.3 %), and aquaculture (21.7 %). High-risk workers had significantly higher PEI (p < 0.001), longer sector employment (p = 0.002), and lower SDNN values (p < 0.001) compared to low-to-moderate risk counterparts. As noted in Section 2.6, these group differences are partially expected given the construction of the OHRI from these same variables.

Table 1.

Sociodemographic and occupational characteristics of the study population
(n = 1,248)

Characteristic	Total (n=1,248)	Low-mod- erate risk (n=881, 70.6 %)	High risk (n=367, 29.4 %)	p-va- lue
Age, years (mean ± SD)	42.7 ± 11.4	41.9 ± 11.1	44.7 ± 11.9	0.001
Male sex, n (%)	728 (58.3)	512 (58.1)	216 (58.9)	0.782
Employment, years (mean ± SD)	12.8 ± 8.3	11.9 ± 7.8	15.1 ± 9.1	0.002
Weekly hours (mean ± SD)	46.2 ± 6.7	45.4 ± 6.2	48.3 ± 7.5	<0.001
PEI (mg/m³·years, mean ± SD)	3.41 ± 2.18	2.87 ± 1.74	4.71 ± 2.63	<0.001

SDNN (ms, mean $\pm$ SD)	38.4 $\pm$ 14.2	42.1 $\pm$ 13.8	29.3 $\pm$ 11.4	<0.001
PM _{2.5} ($\mu\text{g}/\text{m}^3$, mean $\pm$ SD)	18.7 $\pm$ 7.3	16.9 $\pm$ 6.4	23.2 $\pm$ 8.1	<0.001
BMI (kg/m^2 , mean $\pm$ SD)	27.3 $\pm$ 4.6	26.8 $\pm$ 4.4	28.6 $\pm$ 4.9	0.003
Wage % of regional median	87.4 $\pm$ 22.1	89.2 $\pm$ 21.3	82.6 $\pm$ 23.7	0.018

Note. SD – standard deviation; PEI – pesticide exposure index; SDNN – standard deviation of N-N intervals (heart rate variability); BMI – body mass index. p-values from independent-samples t-test (continuous) or chi-square test (categorical). Group differences are partially expected given that the outcome variable (OHRI) is constructed from these predictor variables (see Section 2.6).

3.2. Model Performance

Table 2 summarizes the performance metrics of the three classifiers on the holdout test set ($n = 187$, enterprise-aware split). The ANN model achieved the highest AUC-ROC of 0.872 (95 % CI: 0.831–0.913, bootstrap), with sensitivity of 0.806, specificity of 0.862, and F1 score of 0.801. The ANN significantly outperformed LR-EN ($\Delta\text{AUC} = 0.069$, DeLong $p < 0.001$). The difference between ANN and GBC was not statistically significant ($\Delta\text{AUC} = 0.014$, DeLong $p = 0.214$), indicating comparable discriminatory performance of the two non-linear classifiers for this task. However, the statistical power of the DeLong test to detect a clinically meaningful AUC difference (e.g., $\Delta\text{AUC} = 0.03$) at $\alpha = 0.05$ was limited given the test set size ($n = 187$, 55 positive cases); this non-significant result should therefore be interpreted with caution.

The ANN achieved the lowest Brier score (0.128), and calibration assessment showed a calibration slope of 0.94 (95 % CI: 0.82–1.06) and intercept of -0.03 (95% CI: -0.11 – 0.05), suggesting adequate probabilistic calibration. The Hosmer–Lemeshow test was non-significant ($p = 0.42$). Calibration plots are presented in Supplementary Figure S1.

Table 2.

**Predictive performance of classifiers on holdout test set
($n = 187$, enterprise-aware split)**

Metric	ANN (proposed)	Gradient Boosting	Logistic Regression
AUC-ROC(95%CI)	0.872 (0.831–0.913)	0.858 (0.814–0.902)	0.803 (0.754–0.852)
Sensitivity	0.806	0.782	0.714
Specificity	0.862	0.845	0.822
PPV	0.797	0.770	0.721
NPV	0.869	0.855	0.814

F1 Score	0.801	0.776	0.717
Brier Score	0.128	0.145	0.174
Accuracy	0.843	0.823	0.789

Note. AUC-ROC – area under receiver operating characteristic curve; PPV – positive predictive value; NPV – negative predictive value. 95 % CI computed by 2,000-replicate stratified bootstrap. ANN significantly outperformed Logistic Regression (DeLong $p < 0.001$). The difference between ANN and Gradient Boosting was not statistically significant (DeLong $p = 0.214$). Classification threshold optimized by Youden index on the validation set (threshold = 0.47).

Sensitivity analysis (leave-one-enterprise-out). Leave-one-enterprise-out cross-validation yielded a mean AUC-ROC of 0.854 (range: 0.791–0.912 across the 14 folds), supporting the robustness of the model to enterprise-level variation. [Note: The following figures should be prepared and inserted: Figure 1 – ROC curves for all three classifiers on a single plot; Figure 2 – Calibration plot (currently Supplementary Figure S1, should be promoted to main text); Figure 3 – Bar chart of permutation importance scores for all 22 predictors; Figure 4 – SHAP beeswarm plot for the top 10 predictors (pending SHAP analysis).]

Table 2a.

Confusion matrix for the ANN model on holdout test set (n = 187)

	Predicted High Risk	Predicted Low-Moderate Risk
Actual High Risk (n = 55)	44 (TP)	11 (FN)
Actual Low-Moderate Risk (n = 132)	18 (FP)	114 (TN)

3.3. Predictor Importance

Table 3 presents permutation importance scores for the top 10 of the 22 predictors in the ANN model. The top five predictors by mean importance score were: cumulative pesticide exposure index (PEI, mean importance = 0.083), HRV SDNN (0.071), ambient PM_{2.5} (0.063), subjective labor intensity score (0.054), and years of employment (0.049). These five variables collectively accounted for 53.6 % of the sum of all 22 permutation importance values. Among domain groups, the ecological domain (EES predictors) contributed 41.2 % of total importance, the physiological domain (PSS predictors) 38.9 %, and the socio-occupational domain (SLSS predictors) 19.9 %.

It should be noted that because the OHRI outcome is itself constructed from these predictor domains (see Section 2.2), the observed importance ranking partially reflects the domain weights embedded in the OHRI formula. The

correlation between permutation importance and AHP-derived domain weights should therefore be interpreted as a consistency check on the model's approximation quality rather than as independent evidence of the predictive value of individual risk factors.

Table 3.

Top-10 predictors by permutation importance in the ANN model

Rank	Predictor	Domain	Mean Importance ($\pm$ SD)	Pearson r with OHRI
1	Pesticide Exposure Index (PEI)	Ecological	0.083 ± 0.009	0.61^{***}
2	HRV SDNN (ms)	Physiological	0.071 ± 0.008	-0.57^{***}
3	$PM_{2.5}$ ($\mu\text{g}/\text{m}^3$)	Ecological	0.063 ± 0.007	0.54^{***}
4	Subjective labor intensity (score)	Socio-occupational	0.054 ± 0.006	0.48^{***}
5	Employment, years	Socio-occupational	0.049 ± 0.006	0.45^{***}
6	Systolic BP (mmHg)	Physiological	0.043 ± 0.005	0.43^{***}
7	HMCI (Z_c index)	Ecological	0.041 ± 0.005	0.41^{***}
8	Shift irregularity index	Socio-occupational	0.038 ± 0.005	0.39^{***}
9	BMI (kg/m^2)	Physiological	0.036 ± 0.004	0.37^{***}
10	NO_2 ($\mu\text{g}/\text{m}^3$)	Ecological	0.034 ± 0.004	0.35^{***}

Note. OHRI – Occupational Health Risk Index; HRV SDNN – standard deviation of N-N intervals; HMCI – heavy metal contamination index (Z_c summation index according to Saet et al., 1990); BP – blood pressure; $^{***} p < 0.001$. Pearson correlations with OHRI are partially structural given that OHRI is constructed from these variables (see Section 2.2).

4. Discussion

The present study demonstrates that a feedforward ANN can classify agricultural workers into high-risk and low-to-moderate-risk categories, defined by an expert-constructed composite OHRI, with good discriminatory accuracy (AUC-ROC = 0.872 in enterprise-aware holdout validation). The ANN significantly outperformed logistic regression but did not achieve statistically significant superiority over gradient boosting (XGBoost), consistent with the broader machine learning literature – including applications in construction safety [26] and the petrochemical industry [27] – showing that ensemble tree-based methods – including random forests [17] and gradient boosting [10] – and neural networks often perform comparably on structured tabular data of moderate dimensionality [11].

4.1. Interpretation of Predictor Importance

The prominence of the pesticide exposure index as the most important predictor is consistent with ecological and toxicological evidence on organophosphate and pyrethroid effects on cardiovascular and neurological systems [12]. The second-ranked position of HRV SDNN – a validated physiological marker of autonomic nervous system regulation [13] – and its inverse correlation with OHRI ($r = -0.57$) are consistent with the hypothesis that reduced heart rate variability reflects impaired vagal tone associated with cumulative occupational burden. However, the cross-sectional design precludes temporal or causal inference: it cannot be established from these data whether reduced SDNN precedes or follows the development of elevated occupational health risk.

The contribution of $PM_{2.5}$ as the third-ranked predictor may reflect both agricultural dust generated by field operations and machinery and industrial or transport emissions from the St. Petersburg conurbation, consistent with evidence that outdoor air pollution sources contribute substantially to health burden in urbanized regions [14]. However, the present study did not include meteorological trajectory analysis or source apportionment, and this interpretation remains speculative.

Importantly, because the OHRI outcome is itself a weighted function of the predictor domains, the observed importance ranking should be interpreted as demonstrating that the ANN successfully approximates the expert-defined OHRI structure, rather than as independent evidence of biological dose–response relationships.

4.2. Practical Utility

The practical value of the proposed model lies not in the discovery of novel risk associations, but in the automation of a multi-step expert evaluation process. Currently, OHRI computation requires manual integration of data from three administrative systems, domain-score calculation, and application of AHP-derived weights – a process that demands specialist knowledge and time. The trained ANN model can perform this classification instantaneously from raw input data, facilitating integration into digital workplace management systems. From a practical standpoint, the model may enable: (a) real-time risk flagging during periodic medical examinations, (b) batch screening of enterprise workforces, and (c) identification of subgroups warranting priority occupational health interventions. The socio-occupational domain – represented by subjective labor intensity, shift irregularity, and employment duration – contributed approximately 20 % of total predictor importance. While smaller than the

ecological and physiological domains, this contribution is non-negligible and provides potentially actionable targets for enterprise-level interventions: regularization of shift schedules (consistent with evidence linking atypical work hours to metabolic syndrome [16]), reduction of overtime prevalence, and monitoring of labor intensity.

4.3. Methodological Considerations

An important methodological contribution of this work is the operationalization of the OHRI as a domain-weighted composite index calibrated through AHP expert elicitation [20]. The AHP consistency ratio of 0.04 indicates acceptable agreement among the expert panel. The composite approach provides a transparent and reproducible basis for aggregating heterogeneous indicators into a single risk stratum, addressing a recognized gap in existing occupational risk classification systems that typically treat domains independently [7]. The AHP weighting can be revised as new expert consensus emerges.

The enterprise-aware validation strategy addresses a potential source of data leakage that is often overlooked in occupational health ML studies: when environmental exposures are measured at the group (enterprise) level, naive random splitting can inflate model performance by placing workers from the same enterprise in both training and test sets. The leave-one-enterprise-out analysis (mean AUC 0.854) confirmed that the model generalizes beyond the specific enterprises in the training set, although performance varied across folds (range: 0.791–0.912), likely reflecting differences in enterprise size and risk profile.

4.4. Limitations

The study has several important limitations that should be considered when interpreting the results.

First, and most critically, the OHRI outcome variable is an author-constructed composite index derived from the same data domains that serve as model predictors. This creates a degree of circularity: the ANN model is not predicting an independent clinical outcome (such as diagnosed occupational disease, sick leave, or loss of work capacity), but is rather learning to approximate a deterministic function of its own inputs. Consequently, the reported AUC values should not be interpreted as evidence of clinical predictive validity. Future studies should validate the model against hard clinical endpoints, such as incident occupational disease diagnoses confirmed through the regional center for occupational pathology.

Second, the cross-sectional design precludes causal inference and limits the temporal interpretation of findings. Terms such as “risk prediction” are used in

the statistical (classification) sense, not in the sense of forecasting future disease onset. Prospective cohort studies with longitudinal follow-up are needed to establish whether OHRI classifications at baseline predict subsequent health deterioration.

Third, ecological exposure variables ($PM_{2.5}$, SO_2 , NO_2 , noise) were assigned at the enterprise level from the nearest monitoring station, introducing potential ecological fallacy. Workers within the same enterprise may experience substantially different microenvironments (e.g., greenhouse vs. open-field workers), and this within-enterprise variability is not captured by the model.

Fourth, the study period (2021–2025) overlaps with the COVID-19 pandemic, which may have affected: (a) the completeness and quality of PME data during 2021–2022, (b) workers' health status (post-COVID cardiovascular complications), (c) labor conditions (quarantine restrictions, changes in employment patterns), and (d) environmental monitoring data (reduced industrial emissions in 2021). A sensitivity analysis excluding data from 2021–2022 ($n = 814$) yielded AUC-ROC = 0.865 (95 % CI: 0.818–0.912), suggesting that the pandemic period did not substantially alter the main findings. Additionally, the study population, while representative of the Leningrad Region, may not generalize to agro-industrial workers in other Russian periurban regions with distinct ecological or agricultural profiles. External validation in geographically distinct populations is required.

Fifth, self-reported socio-occupational variables (labor intensity, job satisfaction, perceived employment stability) are subject to social desirability bias, particularly in enterprises where workers may associate high reported intensity with job insecurity. Objective measures (e.g., wearable accelerometry for labor intensity, administrative payroll records for overtime) could strengthen future iterations of the model. Furthermore, complete-case analysis may introduce selection bias if missingness is non-random, and the model's interpretability is currently limited to permutation importance analysis; future work should incorporate SHAP (SHapley Additive exPlanations) values and partial dependence plots to elucidate the direction, magnitude, and non-linearity of individual predictor effects.

4.5. Comparison with Existing Literature

The AUC-ROC of 0.872 achieved by the ANN model in this study is broadly consistent with discriminatory performance reported in comparable machine learning applications in occupational health. Sarkar and Maiti [11] reported AUC values in the range of 0.78–0.89 for ML-based classification of occupational ac-

cident severity in industrial settings. In agricultural health specifically, studies employing gradient boosting and random forest classifiers for pesticide exposure risk have reported AUC values between 0.81 and 0.91 [21, 22]. However, direct comparison is complicated by the fact that most prior studies predicted hard clinical outcomes (e.g., diagnosed occupational disease, injury severity), whereas the present study classifies an internally constructed composite index. The observed AUC of 0.872 therefore reflects the model's ability to approximate the expert-defined OHRI formula rather than its clinical predictive validity, and likely represents an upper bound of the performance that would be observed against independent clinical endpoints. A naïve benchmark - direct application of the OHRI formula followed by threshold classification - would yield perfect classification (AUC = 1.0) if applied to the same data. That the ANN achieves $AUC < 1.0$ reflects the fact that it learns from raw (unstandardized) input features and must implicitly recover the z-scoring, domain aggregation, and min-max normalization embedded in the OHRI construction. Future studies should benchmark ANN performance against simple formula-based classification to quantify the added value of the machine learning approach more transparently.

The enterprise-aware group splitting strategy employed in this study addresses a methodological gap identified in recent reviews of ML applications in occupational health [23]. Many studies in this domain ignore the clustered nature of workplace data, leading to overly optimistic performance estimates. The leave-one-enterprise-out cross-validation (mean AUC 0.854) provides a more conservative and realistic estimate of generalization performance.

5. Conclusion

This study demonstrates that a feedforward artificial neural network can classify an expert-constructed composite occupational health risk index in agricultural workers of the St. Petersburg periurban zone with good discriminatory accuracy (AUC-ROC = 0.872 in enterprise-aware holdout validation). Pesticide exposure, autonomic cardiac regulation (HRV SDNN), and ambient fine particulate matter were identified as the leading contributors to risk classification, while socio-occupational variables – particularly labor intensity and shift irregularity – contributed independently. The ANN performed comparably to gradient boosting and significantly outperformed logistic regression.

The proposed framework offers a practical tool for automated risk stratification in agro-industrial enterprises, potentially reducing the reliance on manual expert evaluation for individual worker assessment. However, it must be emphasized that the current model classifies an internally constructed composite

index rather than an independent clinical outcome, which fundamentally limits its clinical interpretability and distinguishes its reported accuracy from clinical predictive validity. Future research should focus on: (i) prospective validation of OHRI against incident occupational disease endpoints; (ii) longitudinal tracking of OHRI trajectories following workplace interventions; (iii) integration of SHAP-based model interpretability [24] and feature attribution methods [30] for clinical transparency; (iv) application of proactive risk prediction approaches using longitudinal data [29]; (v) comparative benchmarking against naïve formula-based classification; and (vi) extension to other agricultural regions with comparable agro-ecological profiles.

Conflict of interest information. The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Sponsorship information. This research was funded by the Russian Science Foundation, project No. 25-28-01469 “Neural Network Solutions for Managing Social and Labor Relations in the Digital Economy of Megacities.” The authors are grateful to the occupational health services of the participating enterprises and to the Leningrad Regional Environmental Committee (LenKomEco) for providing access to monitoring data.

Data availability statement. The data analyzed in this study were obtained from mandatory periodic medical examination records, workplace sanitary inspection protocols, and regional environmental monitoring databases. Due to the inclusion of personal health information protected under Russian Federal Law No. 152-FZ “On Personal Data,” the individual-level dataset cannot be made publicly available. Aggregated summary statistics and the correlation matrices are available from the corresponding author upon reasonable request, subject to institutional data protection review.

References

1. Rosstat (Federal State Statistics Service). (2024). *Agriculture in Russia* [Statistical collection]. Moscow: Rosstat.
2. Kim, K.-H., Kabir, E., & Jahan, S. A. (2017). Exposure to pesticides and the associated human health effects. *Science of The Total Environment*, 575, 525–535. <https://doi.org/10.1016/j.scitotenv.2016.09.009>. EDN: <https://elibrary.ru/YVWCFZ>
3. Osborne, A., Blake, C., Fullen, B. M., Meredith, D., Phelan, J., McNamara, J., & Cunningham, C. (2012). Prevalence of musculoskeletal disorders among farmers: A

- systematic review. *American Journal of Industrial Medicine*, 55(2), 143–158. <https://doi.org/10.1002/ajim.21033>
4. Virtanen, M., Heikkilä, K., Jokela, M., et al. (2012). Long working hours and coronary heart disease: A systematic review and meta-analysis. *American Journal of Epidemiology*, 176(7), 586–596. <https://doi.org/10.1093/aje/kws139>
 5. Lunner Kolstrup, C., Kallioniemi, M., Lundqvist, P., Kymäläinen, H.-R., Stallones, L., & Brumby, S. (2013). International perspectives on psychosocial working conditions, mental health, and stress of dairy farm operators. *Journal of Agromedicine*, 18(3), 244–255. <https://doi.org/10.1080/1059924X.2013.796903>
 6. Rospotrebnadzor (Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing). (2023). *On the state of sanitary and epidemiological well-being of the population of the Russian Federation in 2022* [State report]. Moscow: Rospotrebnadzor. 364 pp.
 7. Bukhtiyarov, I. V. (2019). Current status and main directions for preserving and strengthening the health of the working population in Russia. *Russian Journal of Occupational Health and Industrial Ecology*, 59(9), 527–532. <https://doi.org/10.31089/1026-9428-2019-59-9-527-532>. EDN: <https://elibrary.ru/HTKRSY>
 8. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
 9. Badri, A., Boudreau-Trudel, B., & Souissi, A. S. (2018). Occupational health and safety in the industry 4.0 era: A cause for major concern? *Safety Science*, 109, 403–411. <https://doi.org/10.1016/j.ssci.2018.06.012>
 10. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
 11. Sarkar, S., & Maiti, J. (2020). Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. *Safety Science*, 131, 104900. <https://doi.org/10.1016/j.ssci.2020.104900>. EDN: <https://elibrary.ru/KJJKWV>
 12. Mostafalou, S., & Abdollahi, M. (2017). Pesticides: An update of human exposure and toxicity. *Archives of Toxicology*, 91(2), 549–599. <https://doi.org/10.1007/s00204-016-1849-x>. EDN: <https://elibrary.ru/YVVOCH>
 13. Thayer, J. F., Åhs, F., Fredrikson, M., et al. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2), 747–756. <https://doi.org/10.1016/j.neubiorev.2011.11.009>
 14. Lelieveld, J., Evans, J. S., Fnais, M., et al. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367–371. <https://doi.org/10.1038/nature15371>

15. Marmot, M., Allen, J., Bell, R., et al. (2012). WHO European review of social determinants of health and the health divide. *Lancet*, 380(9846), 1011–1029. [https://doi.org/10.1016/S0140-6736\(12\)61228-8](https://doi.org/10.1016/S0140-6736(12)61228-8)
16. Violanti, J. M., Burchfiel, C. M., Hartley, T. A., Mnatsakanova, A., Fededulegn, D., Andrew, M. E., Charles, L. E., & Vila, B. J. (2009). Atypical work hours and metabolic syndrome among police officers. *Archives of Environmental & Occupational Health*, 64(3), 194–201. <https://doi.org/10.1080/19338240903241259>
17. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>. EDN: <https://elibrary.ru/ARROTH>
18. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
19. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
20. Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83–98. <https://doi.org/10.1504/IJSS-CI.2008.017590>
21. Damalas, C. A., & Koutroubas, S. D. (2016). Farmers' exposure to pesticides: Toxicity types and ways of prevention. *Toxics*, 4(1), Article 1. <https://doi.org/10.3390/toxics4010001>
22. Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, 117, 257–262. <https://doi.org/10.1016/j.ssci.2019.04.026>
23. Bertke, S. J., Meyers, A. R., Wurzelbacher, S. J., et al. (2012). Development and evaluation of a naïve Bayesian model for coding causation of workers' compensation claims. *Journal of Safety Research*, 43(5–6), 327–332. <https://doi.org/10.1016/j.jsr.2012.10.012>
24. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, 4765–4774.
25. Izmerov, N. F., Tikhonova, G. I., & Churanova, A. N. (2014). Mortality of the working-age population in Russia and industrialized European countries: Trends over the last two decades. *Annals of the Russian Academy of Medical Sciences*, (7–8), 121–126. <https://doi.org/10.15690/vramn.v69i7-8.1118>. EDN: <https://elibrary.ru/SNHYEN>
26. Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62, 45–56. <https://doi.org/10.1016/j.autcon.2015.11.001>

27. Cheng, C.-W., Yao, H.-Q., & Wu, T.-C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6), 1269–1278. <https://doi.org/10.1016/j.jlp.2013.07.002>
28. Bezrukova, G. A., Spirin, V. F., & Novikova, T. A. (2021). Current aspects of occupational hearing loss in agricultural workers. *Hygiene and Sanitation*, 100(10), 1109–1114. <https://doi.org/10.47470/0016-9900-2021-100-10-1109-1114>. EDN: <https://elibrary.ru/SPFVOI>
29. Mohammadfam, I., Kalatpour, O., & Gholamizadeh, K. (2020). Quantitative assessment of safety and health risks in HAZMAT road transport using a hybrid approach: A case study in Tehran. *ACS Chemical Health & Safety*, 27(4), 240–250. <https://doi.org/10.1021/acs.chas.0c00018>. EDN: <https://elibrary.ru/DOVKOO>
30. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, 70, 3145–3153.

Список литературы

1. Росстат (Федеральная служба государственной статистики). (2024). *Сельское хозяйство в России* [Статистический сборник]. Москва: Росстат.
2. Kim, K.-H., Kabir, E., & Jahan, S. A. (2017). Exposure to pesticides and the associated human health effects. *Science of The Total Environment*, 575, 525–535. <https://doi.org/10.1016/j.scitotenv.2016.09.009>. EDN: <https://elibrary.ru/YVWCFZ>
3. Osborne, A., Blake, C., Fullen, B. M., Meredith, D., Phelan, J., McNamara, J., & Cunningham, C. (2012). Prevalence of musculoskeletal disorders among farmers: A systematic review. *American Journal of Industrial Medicine*, 55(2), 143–158. <https://doi.org/10.1002/ajim.21033>
4. Virtanen, M., Heikkilä, K., Jokela, M., et al. (2012). Long working hours and coronary heart disease: A systematic review and meta-analysis. *American Journal of Epidemiology*, 176(7), 586–596. <https://doi.org/10.1093/aje/kws139>
5. Lunner Kolstrup, C., Kallioniemi, M., Lundqvist, P., Kymäläinen, H.-R., Stallones, L., & Brumby, S. (2013). International perspectives on psychosocial working conditions, mental health, and stress of dairy farm operators. *Journal of Agromedicine*, 18(3), 244–255. <https://doi.org/10.1080/1059924X.2013.796903>
6. Роспотребнадзор. (2023). *О состоянии санитарно-эпидемиологического благополучия населения Российской Федерации в 2022 году* [Государственный доклад]. Москва: Федеральная служба по надзору за защитой прав потребителей и благополучия человека. 364 с.
7. Бухтияров, И. В. (2019). Современное состояние и основные направления сохранения и укрепления здоровья трудящегося населения России. *Российский*

- журнал гигиены труда и промышленной экологии, 59(9), 527–532. <https://doi.org/10.31089/1026-9428-2019-59-9-527-532>. EDN: <https://elibrary.ru/HTKRSY>
8. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
 9. Badri, A., Boudreau-Trudel, B., & Souissi, A. S. (2018). Occupational health and safety in the industry 4.0 era: A cause for major concern? *Safety Science*, 109, 403–411. <https://doi.org/10.1016/j.ssci.2018.06.012>
 10. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. B *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
 11. Sarkar, S., & Maiti, J. (2020). Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. *Safety Science*, 131, 104900. <https://doi.org/10.1016/j.ssci.2020.104900>. EDN: <https://elibrary.ru/KJJKWV>
 12. Mostafalou, S., & Abdollahi, M. (2017). Pesticides: An update of human exposure and toxicity. *Archives of Toxicology*, 91(2), 549–599. <https://doi.org/10.1007/s00204-016-1849-x>. EDN: <https://elibrary.ru/YVVOCH>
 13. Thayer, J. F., Åhs, F., Fredrikson, M., et al. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2), 747–756. <https://doi.org/10.1016/j.neubiorev.2011.11.009>
 14. Lelieveld, J., Evans, J. S., Fnais, M., et al. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367–371. <https://doi.org/10.1038/nature15371>
 15. Marmot, M., Allen, J., Bell, R., et al. (2012). WHO European review of social determinants of health and the health divide. *Lancet*, 380(9846), 1011–1029. [https://doi.org/10.1016/S0140-6736\(12\)61228-8](https://doi.org/10.1016/S0140-6736(12)61228-8)
 16. Violanti, J. M., Burchfiel, C. M., Hartley, T. A., Mnatsakanova, A., Fededulegn, D., Andrew, M. E., Charles, L. E., & Vila, B. J. (2009). Atypical work hours and metabolic syndrome among police officers. *Archives of Environmental & Occupational Health*, 64(3), 194–201. <https://doi.org/10.1080/19338240903241259>
 17. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>. EDN: <https://elibrary.ru/ARROTH>
 18. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics*, 44(3), 837–845. <https://doi.org/10.2307/2531595>
 19. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

20. Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83–98. <https://doi.org/10.1504/IJSS-CI.2008.017590>
21. Damalas, C. A., & Koutroubas, S. D. (2016). Farmers' exposure to pesticides: Toxicity types and ways of prevention. *Toxics*, 4(1), Article 1. <https://doi.org/10.3390/toxics4010001>
22. Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, 117, 257–262. <https://doi.org/10.1016/j.ssci.2019.04.026>
23. Bertke, S. J., Meyers, A. R., Wurzelbacher, S. J., et al. (2012). Development and evaluation of a naïve Bayesian model for coding causation of workers' compensation claims. *Journal of Safety Research*, 43(5–6), 327–332. <https://doi.org/10.1016/j.jsr.2012.10.012>
24. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. B *Advances in Neural Information Processing Systems*, 30, 4765–4774.
25. Измеров, Н. Ф., Тихонова, Г. И., & Чуранова, А. Н. (2014). Смертность трудоспособного населения в России и промышленно развитых странах Европы: тенденции последних двух десятилетий. *Вестник Российской академии медицинских наук*, (7–8), 121–126. <https://doi.org/10.15690/vramn.v69i7-8.1118>. EDN: <https://elibrary.ru/SNHYEN>
26. Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62, 45–56. <https://doi.org/10.1016/j.autcon.2015.11.001>
27. Cheng, C.-W., Yao, H.-Q., & Wu, T.-C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6), 1269–1278. <https://doi.org/10.1016/j.jlp.2013.07.002>
28. Безрукова, Г. А., Спирин, В. Ф., & Новикова, Т. А. (2021). Актуальные аспекты профессиональной потери слуха у сельскохозяйственных рабочих. *Гигиена и санитария*, 100(10), 1109–1114. <https://doi.org/10.47470/0016-9900-2021-100-10-1109-1114>. EDN: <https://elibrary.ru/SPFVOI>
29. Mohammadfam, I., Kalatpour, O., & Gholamizadeh, K. (2020). Quantitative assessment of safety and health risks in HAZMAT road transport using a hybrid approach: A case study in Tehran. *ACS Chemical Health & Safety*, 27(4), 240–250. <https://doi.org/10.1021/acs.chas.0c00018>. EDN: <https://elibrary.ru/DOVKOO>
30. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. B *Proceedings of the 34th International Conference on Machine Learning*, 70, 3145–3153.

DATA ABOUT THE AUTHORS

Li Yin, PhD, Associate Professor, School of Management

Harbin Institute of Technology

Harbin, China

dr.liyin@hit.edu.cn

ORCID: <https://orcid.org/0000-0001-6474-9289>

Olga T. Ergunova, PhD in Economics, Associate Professor, Associate Professor, Graduate School of Industrial Management

Peter the Great St. Petersburg Polytechnic University

29, Politekhnikeskaya Str., Saint Petersburg, 195251, Russian Federation

ergunova-olga@yandex.ru

ORCID: <https://orcid.org/0000-0002-1714-7784>

Aleksandr B. Titov, PhD in Economics, Associate Professor, Associate Professor at the Graduate School of Industrial Management; Associate Professor at the Department of Business Informatics

Peter the Great St. Petersburg Polytechnic University; Financial University under the Government of the Russian Federation (St. Petersburg branch)

29, Politekhnikeskaya Str., Saint Petersburg, 195251, Russian Federation; 15-17A, Syezhinskaya Str., St. Petersburg, 197198, Russian Federation

titov_ab@spbstu.ru

ORCID: <https://orcid.org/0000-0002-9014-2541>

Andrey G. Somov, PhD in Economics, Senior Lecturer at the Graduate School of Industrial Management

Peter the Great St. Petersburg Polytechnic University

29, Politekhnikeskaya Str., Saint Petersburg, 195251, Russian Federation

somovspb@yandex.ru

ORCID: <https://orcid.org/0009-0004-2592-9198>

ДАННЫЕ ОБ АВТОРАХ

Ли Инь, PhD, доцент, Школа менеджмента

Харбинский технологический институт

Харбин, Китай

dr.liyin@hit.edu.cn

Ергунова Ольга Титовна, кандидат экономических наук, доцент, доцент Высшей школы производственного менеджмента

*Санкт-Петербургский политехнический университет Петра Великого
ул. Политехническая, 29, г. Санкт-Петербург, 195251, Российская Федерация
ergunova-olga@yandex.ru*

Титов Александр Борисович, кандидат экономических наук, доцент, доцент Высшей школы производственного менеджмента; доцент кафедры «Бизнес-информатика»

*Санкт-Петербургский политехнический университет Петра Великого; Финансовый университет при Правительстве РФ (Санкт-Петербургский филиал)
ул. Политехническая, 29, г. Санкт-Петербург, 195251, Российская Федерация
ул. Съезжинская, 15-17А, г. Санкт-Петербург, 197198, Российская Федерация
titov_ab@spbstu.ru*

Сомов Андрей Георгиевич, кандидат экономических наук, старший преподаватель Высшей школы производственного менеджмента

*Санкт-Петербургский политехнический университет Петра Великого
ул. Политехническая, 29, г. Санкт-Петербург, 195251, Российская Федерация
somovspb@yandex.ru*

Поступила 11.02.2026

После рецензирования 21.03.2026

Принята 27.03.2026

Received 11.02.2026

Revised 21.03.2026

Accepted 27.03.2026