Scientific Research Publishing

Health, 2025, 17(5), 579-593
https://www.scirp.org/journal/health
ISSN Online: 1949-5005
ISSN Print: 1949-4998

Occupational Diseases Risk Prediction by Neural Networks

Paolo Montanari

Department of Medicine, Epidemiology, Occupational and Environmental Hygiene, National Institute for Insurance against Accidents at Work (INAIL), Rome, Italy
Email: p.montanari@inail.it

How to cite this paper: Montanari, P. (2025) Occupational Diseases Risk Prediction by Neural Networks. *Health*, **17**, 579-593.
https://doi.org/10.4236/health.2025.175037

Received: January 31, 2025
Accepted: May 27, 2025
Published: May 30, 2025

Abstract

This study explores the use of neural networks for occupational disease risk prediction based on worker and workplace characteristics. The goal is to develop a tool to assist occupational physicians in monitoring workers. Using a dataset from the Italian MalProf National Surveillance System (2019-2023), an ensemble of one-vs-all classifiers is trained to identify six prevalent disease classes. Performance is evaluated using accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The results indicate promising performance. The specificity values for all six disease classes under study exceed 0.920 on average over 10 runs, and for five out of six classes, they surpass 0.967. Regarding sensitivity, the performance is positive (average over 10 runs greater than 0.920) for all classes, except for "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb", which performs less effectively (average over 10 runs = 0.655). Future research could focus on optimizing neural network architectures, applying oversampling techniques for underrepresented classes, and analyzing misclassifications.

Keywords

Occupational Disease, Machine Learning, Supervised Learning, Neural Network, Deep Learning

1. Introduction

Employee health is a major concern for companies worldwide, not only for legal and ethical reasons but also for economic ones, as each case of occupational disease entails significant direct and indirect costs. The World Health Organization (WHO) also promotes initiatives aimed at improving workers' physical, mental, and social well-being. Many countries are increasing their healthcare expendi-

tures, not only due to rising life expectancy but also because of the extension of working life and the consequent prolonged exposure to risk factors, even at an advanced age. Healthcare costs could be reduced by expanding the use of machine learning algorithms in occupational medicine. This study explores the use of neural networks on a database of occupational diseases analyzed between 2019 and 2023 by local health authorities participating in the Italian MalProf National Surveillance System. Compared to the archive used for compensation purposes, MalProf contains fewer records but provides more detailed information, particularly regarding workers' employment history. In 2019, significant modifications were made to the record structure, including the adoption of updated classification systems (e.g., ICD 10 for disease classification) and the addition of the exposure agent variable. For this reason, the two sections of the archive (pre- and post-2019) cannot be analysed jointly.

In recent years, data mining and machine learning techniques have been applied to many problems in occupational medicine. A decision support system for employee healthcare was developed in [1]; in [2], clustering techniques were applied to medical data to predict the likelihood of diseases; in [3], Artificial Neural Networks were used to forecast the incidence of occupational diseases; in [4], an Artificial Neural Network was employed to predict workers' pneumoconiosis in an iron and steel company; in [5], a fuzzy system based on the Mamdani inference model was proposed to support the diagnosis of musculoskeletal disorders; in [6], a neuro-fuzzy network was employed to forecast absenteeism at work due to either short-term or long-term diseases; in [7], an Artificial Neural Network was used to identify and classify different levels of pneumoconiosis risk in coal miners with varying work histories; in [8], three machine learning techniques (Naïve Bayes, Decision Tree and Artificial Neural Network) were applied to forecast heart conditions in coal miners. Works [9] and [10] proposed three variants of an unsupervised classification system based on clustering, incorporating genetic optimization as an automatic feature selection engine, to predict the likelihood of contracting a disease based on worker and workplace characteristics. In [11], a study compared different machine learning techniques, including SVM, for predicting the risk of occupational diseases. The studies in [9] [10], and [11] analysed data from the 1999-2009 period of the same MalProf system that is the subject of this study. Compared to previous research, this study, like [9] [10], and [11], simultaneously analyses multiple disease classes but employs neural networks and utilizes a database with more recent data. Analysing multiple disease classes simultaneously is essential for developing a unified tool to support occupational physicians in monitoring workers across different work sectors. In the Discussion section, the results of [10] and [11] are compared with those of the present study.

2. Materials and Methods

2.1. Dataset Description

The dataset used in this study derives from the data entered from 2019 to 2023 by

the Italian Local Health Authorities adhering to the MalProf Surveillance system. In 2019, the record structure changed, through the updating of some classification systems and the addition of some variables including "exposure agent". As a result of the changes, the data from years prior to 2019 are not consistent with those of more recent years. The MalProf dataset contains data on reports of diseases of suspected occupational origin. For each report, the following data are stored: disease (coded with the International Classification of Diseases—ICD 10), age, gender and, for each work period, duration of the work period, sector of Economic Activity of the company (Ateco), worker's activity, exposure agent. The connection between disease and each work period is graded according to a four-values scale: "highly probable", "probable", "improbable", "highly improbable". In this study, only work periods positively ("highly probable" or "probable") linked to disease reports are considered.

The number of records is higher than the number of disease reports because each disease report can be associated with more than one period (Figure 1).

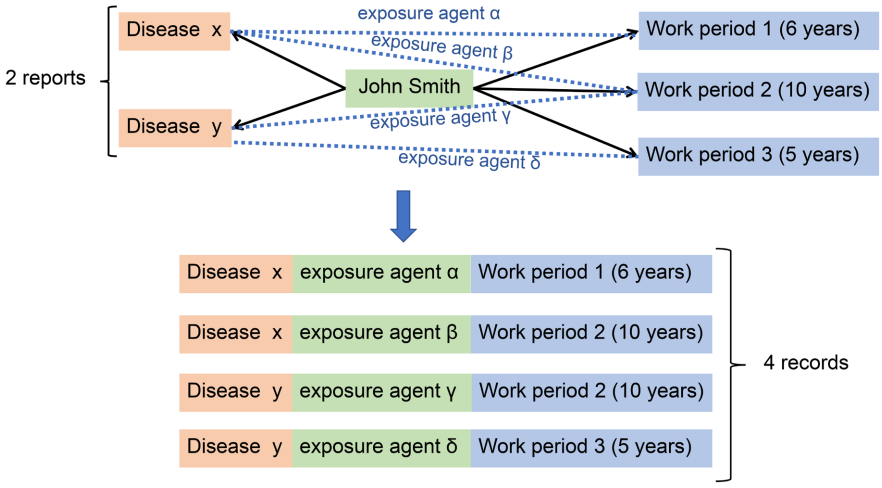


Figure 1. Example of the association between workers' diseases and work periods.

The descriptive variables are listed in Table 1. The sector of economic activity of the company is based on the Italian version of the NACE classification system. NACE (Nomenclature des Activités économiques dans la Communauté Européenne) is a European industry standard classification system similar in function to the Standard Industry Classification (SIC) and the North American Industry Classification System (NAICS), used for classifying business activities. The classification of workers' activities is carried out by ISTAT (the Italian National Institute of Statistics) and results from a process of aligning Italian specificities with the International Standard Classification of Occupations—ISCO-08. The classification of protective agents includes 19 items. The most frequent ones are: "Biomechanical overload of the upper and/or lower limbs" (44.18%), "Manual handling of loads" (27.91%), "Noise" (7.79%), "Asbestos" (4.94%), "Postural risks" (4.18%), and "Whole-body vibrations" (2.77%).

Table 1. Features.

Code	Meaning
x_1	Region of receipt of the report
x_2	Worker's gender
x_3	Sector of economic activity of the company
x_4	Worker's activity
x_5	Exposure agent
x_6	Worker's age at the time of reporting
x_7	Duration of work period (years)

After removing records with missing values in at least one of the descriptive variables, the dataset contains 45,846 records. The ICD 10 codes of the diseases were grouped into homogeneous classes by an occupational physician. Table 2 presents the distribution of disease classes. Items with a frequency below 0.9% are grouped into the "Other" category. Table 2(a) to Table 2(f) show the details of disease distributions within each class.

Table 2. Distribution of disease classes in the dataset. (a) Distribution of diseases in the "Musculoskeletal Diseases (excluding Spinal Diseases)" class; (b) Distribution of diseases in the "Spinal Diseases" class; (c) Distribution of diseases in the "Mononeuropathies of Upper Limb" class; (d) Distribution of diseases in the "Ear Disorders (including Hearing Loss)" class; (e) Distribution of diseases in the "Mesothelioma, Asbestosis and Pleural Plaque" class; (f) Distribution of diseases in the "Infectious Diseases" class.

Disease class	N	%
Musculoskeletal Diseases (excluding Spinal Diseases)	18,986	41.41
Spinal Diseases	13,525	29.50
Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb	4561	9.95
Ear Disorders (including Hearing Loss)	3592	7.83
Mesothelioma, Asbestosis and Pleural Plaque	2130	4.65
Infectious Diseases	413	0.90
Other	2639	5.76
Total	**45,846**	**100.00**

Continued

(a)			
ICD 10	**Disease**	**N**	**%**
M00-M25*	Arthropathies	2346	12.36
M60-M79*	Soft tissue disorders	16,611	87.49
M80-M94*	Osteopathies and chondropathies	21	0.11
M96.0	Pseudarthrosis after fusion or arthrodesis	2	0.01
M99.8, M99.9	Other and unspecified biomechanical lesions	6	0.03
	Total	**18,986**	**100.00**
(b)			
ICD 10	**Disease**	**N**	**%**
M40-M54*	Dorsopathies	13,523	99.99
M99.5	Intervertebral disc stenosis of neural canal	2	0.01
	Total	**13,525**	**100.00**
(c)			
ICD 10	**Disease**	**N**	**%**
G56.0	Carpal tunnel syndrome	4332	94.98
G56.1, G56.2, G56.3, G56.8, G56.9	Other mononeuropathies of upper limb	229	5.02
	Total	**4561**	**100.00**
(d)			
ICD 10	**Disease**	**N**	**%**
H80-H83*	Diseases of inner ear	2020	56.24
H90-H95*	Other disorders of ear	1572	43.76
	Total	**3592**	**100.00**
(e)			
ICD 10	**Disease**	**N**	**%**
C45*	Mesothelioma	1253	58.83
J61*	Pneumoconiosis due to asbestos and other mineral fibres	324	15.21
J92*	Pleural plaque	553	25.96
	Total	**2130**	**100.00**

Continued

(f)			
ICD 10	**Disease**	**N**	**%**
A15-A19*	Tuberculosis	6	1.45
A65-A69*	Other spirochaetal diseases	1	0.24
B15-B19*	Viral hepatitis	1	0.24
B25-B34*	Other viral diseases	6	1.45
B35-B49*	Mycoses	4	0.97
B85-B89*	Pediculosis, acariasis and other infestations	2	0.48
J09-J18*	Influenza and pneumonia	13	3.15
J20-J22*	Other acute lower respiratory infections	1	0.24
U07.1, U07.2	COVID-19	379	91.77
	Total	**413**	**100.00**

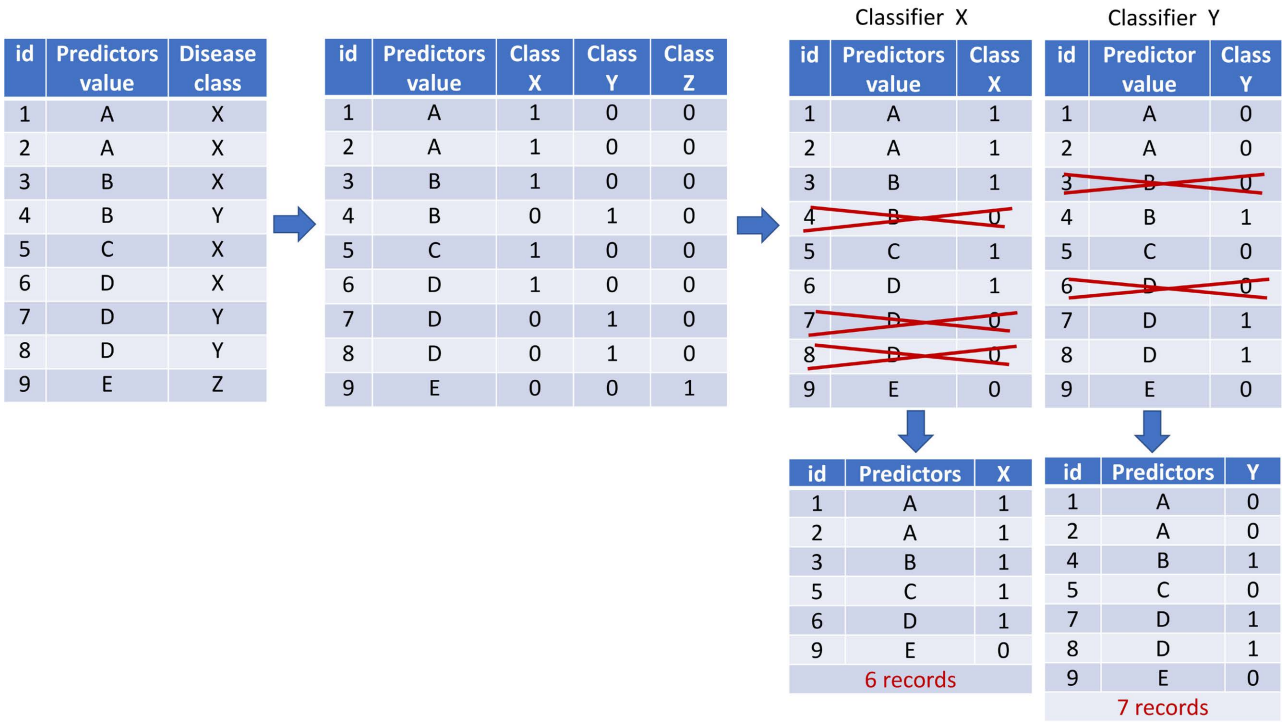
*Including all subcategories.

2.2. Preprocessing

The preprocessing phase was carried out in several steps. First, for each of the 33 values of the variable "disease class", a corresponding dummy variable was created, with a value of 1 assigned if the record was associated with that disease class, and 0 otherwise (as indicated by the leftmost horizontal arrow in Figure 2). For each record, only one dummy column is set to 1, so the sum of the 1 values across the 33 dummy columns equals the total number of records (45,846). Next, to reduce data dispersion, the categories of some descriptive variables were grouped to form new, internally homogeneous, groups. This resulted in 44 values for the variable x_3 (sector of economic activity of the company), 20 values for x_4 (worker's activity), 8 values for x_6 (age of the worker in years at the time of reporting), and 7 values for x_7 (duration of work period in years).

For each of the six classifiers, the operational dataset consists of the seven descriptive variables (x_1, ⋯, x_7) and the disease class under consideration. The steps for constructing the datasets of the six classifiers from the initial dataset are summarized in Figure 2, where, for graphical reasons, the set of descriptive variables is represented in a single column named "predictors value". Analysis of the datasets revealed the presence of records with identical values for the predictive variables but different values for the disease class (target variable), such as records with IDs 3 and 4. These situations may arise either when the same worker reports two or more diseases or when different workers with the same characteristics (in terms of the considered parameters) report different diseases. To resolve these ambiguities, if a classifier's dataset contains records with identical values for the descriptive variables but different values for the disease class, the records where

the disease class variable is 0 are eliminated (e.g., in classifier X's dataset, records with IDs 4, 7, and 8 are removed, while in classifier Y's dataset, records with IDs 3 and 6 are removed). For classifier Z's dataset, which is not shown in Figure 2, no records need to be removed. This procedure resulted in classifiers trained on datasets of different sizes: 39,153 records for the classifier for the disease class "Musculoskeletal Diseases (excluding Spinal Diseases)", 41,270 for the classifier for the class "Spinal Diseases", 36,041 for the classifier for the class "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb", 45,054 for the classifier for the class "Ear Disorders (including Hearing Loss)", 45,693 for the classifier for the class "Mesothelioma, Asbestosis and Pleural Plaque", and 45,837 for the classifier for the class "Infectious Diseases". Each classifier, described in the following paragraph, uses the seven descriptive variables listed in Table 1 as predictors and the corresponding dummy column for the disease class under consideration as the target.



id	Predictors value	Disease class
1	A	X
2	A	X
3	B	X
4	B	Y
5	C	X
6	D	X
7	D	Y
8	D	Y
9	E	Z

id	Predictors value	Class X	Class Y	Class Z
1	A	1	0	0
2	A	1	0	0
3	B	1	0	0
4	B	0	1	0
5	C	1	0	0
6	D	1	0	0
7	D	0	1	0
8	D	0	1	0
9	E	0	0	1

id	Predictors value	Class X
1	A	1
2	A	1
3	B	1
4	B	0
5	C	1
6	D	1
7	D	0
8	D	0
9	E	0

id	Predictors	X
1	A	1
2	A	1
3	B	1
5	C	1
6	D	1
9	E	0

id	Predictor value	Class Y
1	A	0
2	A	0
3	B	0
4	B	1
5	C	0
6	D	0
7	D	1
8	D	1
9	E	0

id	Predictors	Y
1	A	0
2	A	0
4	B	1
5	C	0
7	D	1
8	D	1
9	E	0

Figure 2. Steps to eliminate ambiguity.

2.3. Data Analysis

The proposed approach consists in defining the overall classifier as an ensemble of specific *One-vs-All* classifiers, each trained to recognize patterns of descriptive variables characterizing each disease class. This choice offers great flexibility, allowing a customized solution for each class. The experimentation with multi-label classification, which could reveal relationships between disease classes, will be addressed in subsequent studies. In this study, only the six most frequent disease classes were considered (Table 2).

For each classifier, there is only one binary target variable, which takes the value 1 (positive) for the records of the investigated disease class and 0 (negative) for the records of all other disease classes. Therefore, all records reporting other occupational diseases, with respect to the one under examination, were considered negative. All predictor variables were treated as categorical and processed using the *one-hot encoding* technique, which creates as many binary nodes as there are categories. For each variable, one and only one node is set to 1, while the others are set to 0. In total, the input nodes of the neural network are 117. The output neuron represents the disease class, and the desired output is 1 if the record corresponds to the investigated disease class, and 0 otherwise.

The MATLAB environment was used to implement the procedures. To define the structure of the neural network and the hyperparameters, several experiments were conducted to find the best balance between learning ability (evaluated on the training datasets) and generalization ability (evaluated on the test datasets). For five of the six disease classes, a few experiments were sufficient to define a baseline network with two fully connected hidden layers containing 64 and 32 neurons, respectively, and a single output neuron (**Figure 3**).

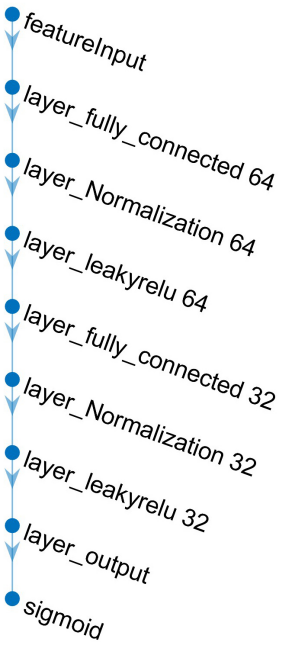


Figure 3. Baseline neural network structure.

For the "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb" class, hyperparameters tuning was performed using a systematic approach such as grid search. **Figure 4** shows the performance on the test set, measured as the average of the mean values over 10 runs for sensitivity, specificity, PPV, and NPV across three neural network architectures, varying in the maximum number of epochs. The light blue line represents the performance of the architecture with two hidden layers containing 64 and 32 nodes, respectively. The orange line rep-

resents the performance of the architecture with three hidden layers containing 128, 64, and 32 nodes, respectively. The grey line represents the performance of the architecture with four hidden layers containing 256, 128, 64, and 32 nodes, respectively. The highest performance is achieved by the architecture with four hidden layers at 40 epochs. The *leakyReLU* activation function was used for the neurons of the hidden layers, and the *sigmoid* function for the output neuron. A normalization layer was inserted downstream of each hidden layer. Furthermore, the following hyperparameters were set: the *Adam* optimizer was used, the maximum number of epochs was set to 20 (except for the "Carpal Tunnel Syndrome" class, where it was set to 40), the minibatch size (which defines the size of the subsets into which the training dataset is divided to improve performance during training) was set to 512, and shuffling was enabled. For each classifier, a seed was set for the random number generator, and ten runs were performed. Within each run, the dataset was randomly divided into a training set (75%) and a test set (25%), maintaining the proportion of sick and healthy individuals in each of the two datasets as in the entire dataset. A different seed was used each time to generate the training set, the test set and the initial parameters, making each run an independent test. The *trainnet* function, introduced with the R2023b version of MATLAB and recommended over the previous *trainNetwork*, was used for training, along with the *binary cross-entropy loss function*. The learning rate was kept at a constant value of 0.001 (preset by MATLAB), as experiments with a constant learning rate of 0.01 or a decreasing learning rate starting from 0.1 resulted in worse performance.

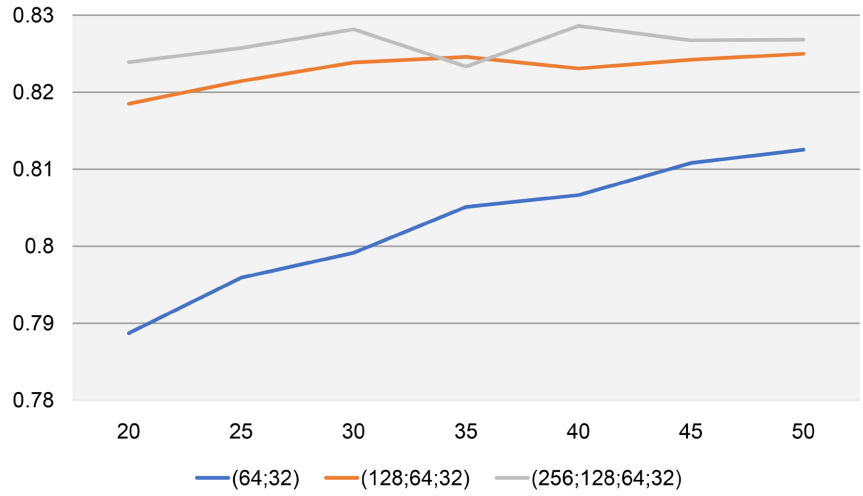


Figure 4. Grid search-based optimization of the neural network architecture and maximum number of epochs for the Carpal Tunnel Syndrome classification task.

3. Results

Five performance measures were considered:

1) ***Accuracy***, defined as the ratio of the number of correctly classified patterns to the total number of patterns.

2) ***Sensitivity***, defined as the ratio of the number of positive patterns correctly classified as such to the total number of positive patterns.

3) ***Specificity***, defined as the ratio of the number of negative patterns correctly classified as such to the total number of negative patterns.

4) ***Positive Predictive Value*** **(PPV)**, defined as the ratio of the number of positive patterns correctly classified as such to the total number of patterns classified as positive (true or false positives).

5) ***Negative Predictive Value*** **(NPV)**, defined as the ratio of negative patterns correctly classified as such to the total number of patterns classified as negative (true or false negatives).

For each disease class, Table 3 shows the mean values and the values corresponding to the best of the ten runs for the five indicators listed above. The best run is the one with the highest Informedness value, defined as $J = Sensitivity + Specificity - 1$. The values on the left were calculated at the end of training on the training set. For quick comparison, the values calculated on the test set are shown on the right and in brackets.

Table 3. Performance over ten runs for the six considered disease classes. Performance on the training set is shown on the left; performance on the test set is shown on the right and in brackets.

Disease class		Performance measure				
		Accuracy	Sensitivity	Specificity	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
Musculoskeletal Diseases (excluding Spinal Diseases)	Best of ten runs[a]	0.956 (0.927)	0.962 (0.920)	0.950 (0.934)	0.948 (0.929)	0.964 (0.925)
	Average	0.953 (0.924)	0.953 (0.922)	0.954 (0.925)	0.951 (0.921)	0.956 (0.927)
Spinal Diseases	Best of ten runs[a]	0.982 (0.966)	0.969 (0.944)	0.988 (0.977)	0.974 (0.952)	0.985 (0.973)
	Average	0.979 (0.962)	0.965 (0.938)	0.986 (0.974)	0.971 (0.947)	0.983 (0.970)
Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb	Best of ten runs[a]	0.999 (0.929)	0.997 (0.694)	0.999 (0.964)	0.993 (0.734)	1.000 (0.956)
	Average	0.997 (0.927)	0.986 (0.655)	0.999 (0.967)	0.990 (0.742)	0.998 (0.951)
Ear Disorders (including Hearing Loss)	Best of ten runs[a]	0.999 (0.999)	0.995 (0.987)	1.000 (1.000)	0.997 (0.998)	1.000 (0.999)
	Average	0.999 (0.998)	0.991 (0.982)	1.000 (1.000)	0.998 (0.997)	0.999 (0.998)
Mesothelioma, Asbestosis and Pleural Plaque	Best of ten runs[a]	0.998 (0.993)	0.997 (0.957)	0.998 (0.995)	0.970 (0.906)	1.000 (0.998)
	Average	0.998 (0.993)	0.992 (0.934)	0.999 (0.996)	0.972 (0.925)	1.000 (0.997)
Infectious Diseases	Best of ten runs[a]	1.000 (1.000)	1.000 (0.981)	1.000 (1.000)	0.997 (0.990)	1.000 (1.000)
	Average	1.000 (0.999)	0.994 (0.952)	1.000 (1.000)	0.997 (0.980)	1.000 (1.000)

[a]Run with the highest Informedness value, defined as $J = Sensitivity + Specificity - 1$.

The results obtained by the six classifiers on the test set are shown in Figure 5. The outcomes of the ten runs are represented using box plots: red horizontal markers indicate the median value; the top and bottom edges of each box correspond to 75th and 25th percentiles, respectively; the whiskers extend to the maximum and minimum values.

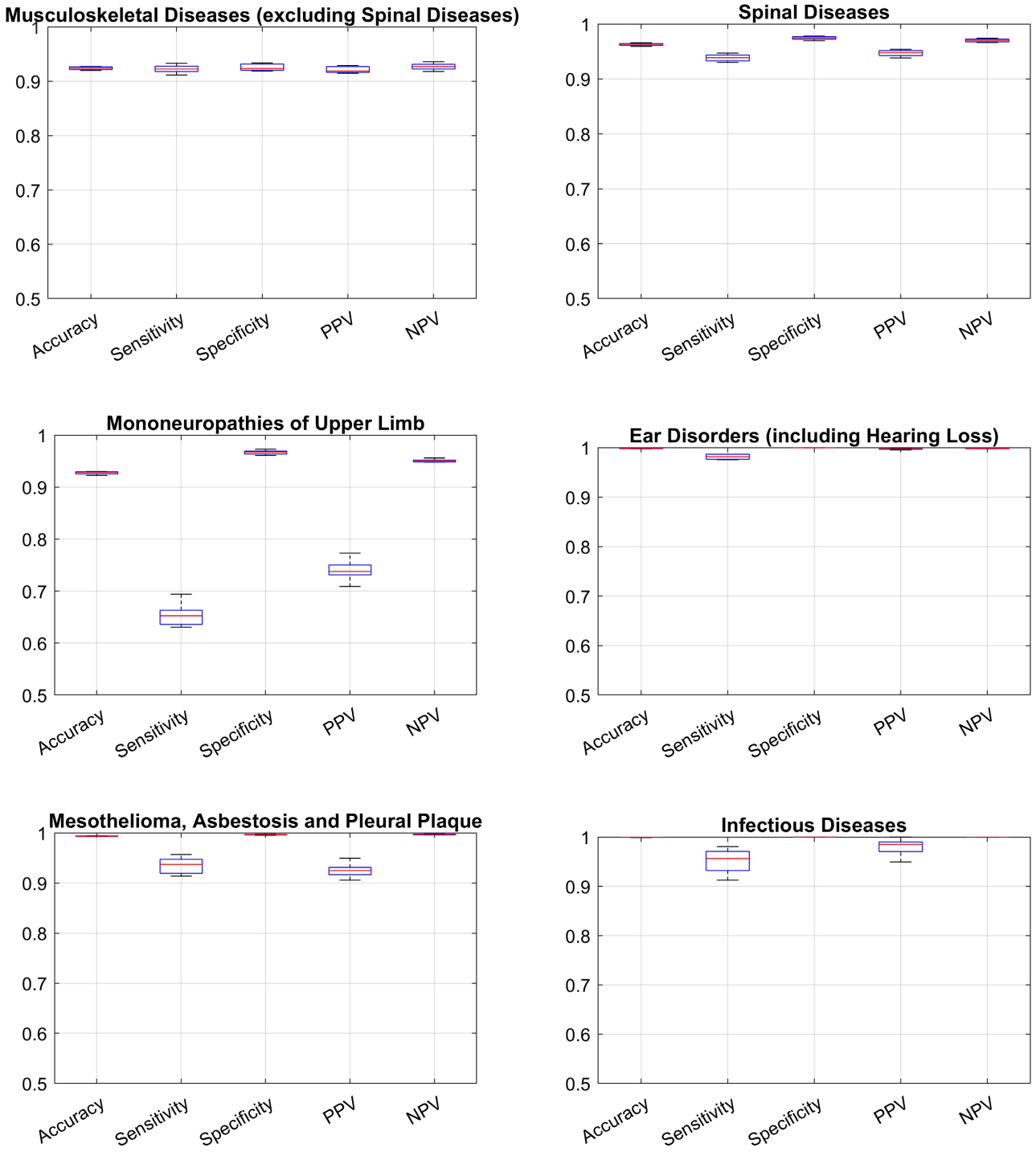


Figure 5. Box plots of performance over ten runs on the test set for the six considered disease classes.

4. Discussion

The proposed approach of building an overall classifier as an ensemble of specific *One-vs-All* classifiers shows encouraging results. As illustrated by the boxplots in Figure 5, five disease classes—"Musculoskeletal Diseases", "Spinal Diseases", "Ear Disorders", "Mesothelioma, Asbestosis and Pleural Plaque", and "Infectious Diseases"—achieve highly positive performances, with mean values across the ten runs exceeding 0.9 for all five indicators, and even higher values in the best run. The only classifier with less brilliant performance is the one dedicated to "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb". Specifically, its average sensitivity is 0.66, while the PPV is 0.74. The robustness of the six classifiers is demonstrated by the low variability of the five indicators across the ten runs. Furthermore, the absence of outliers confirms the stability of the models.

The use of an ensemble of specific One-vs-All classifiers, rather than a multi-label classification approach, has the drawback of not capturing relationships between disease classes. However, it offers the advantage of customizing the neural network architecture and hyperparameters for each disease class individually. In this study, we leveraged this flexibility to optimize the neural network architecture and maximum number of epochs for class "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb". A systematic approach, such as grid search or Bayesian optimization, applied to the remaining five disease classes could potentially enhance classifier performance. Moreover, it could be interesting to experiment with oversampling techniques and customized penalty functions for underrepresented disease classes, such as "Mesothelioma, Asbestosis, and Pleural Plaque" and "Infectious Diseases", both accounting for less than 5% of cases. Misclassifications were not analysed in this study. Future research could examine the predictor variable values associated with misclassifications, particularly for "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb", to identify recurring patterns and potential model improvements. Additionally, a future study could evaluate the importance of each feature to enhance the interpretability of the model and its recommendations, also leveraging knowledge of occupational risk factors.

Table 4 presents a comparison between the results of studies [10] and [11] and those of the present research. Study [10] introduced an unsupervised classification approach based on clustering, utilizing genetic optimization as an automatic feature selection mechanism to estimate the probability of developing a disease based on various worker and workplace characteristics. In [11], researchers explored different machine learning techniques, including SVM, to predict the risk of occupational diseases. Both studies [10] and [11] analysed data from the MalProf system covering the 1999-2009 period, the same system examined in the present research. The table presents the values of the five metrics used for performance evaluation—sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)—along with the Informedness value, defined *as J = Sensitivity + Specificity* − 1. The first column (2025) refers to this study; the second

(2019 SVM – Table 6) and third (2019 SVM – Table 7) columns represent two different implementations of the SVM technique from study [11]; and the fourth (2016 Cluster + GA) refers to study [10]. Only five of the six classes are considered, as the "Infectious Diseases" class was not examined in previous studies. For each disease class, the highest J value is highlighted in bold. Using the J value for comparison, the 2025 model achieves the best performance in four out of the five disease classes, while for the remaining one ("Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb") the results are nearly identical.

Table 4. Comparison of the performance of four models on the test set over ten runs for the considered disease classes. The best performance for each disease class is highlighted in bold, and the Informedness (J = Sensitivity + Specificity − 1) is shown in the last line of each disease class.

Disease Class (ID)		2025	2019 SVM (Table 6)	2019 SVM (Table 7)	2016 cluster + GA
Spinal Diseases	Accuracy	**0.962**	0.81	0.81	-
	Sensitivity	**0.938**	0.04	0.08	0.747
	Specificity	0.974	**0.99**	0.98	0.853
	PPV	**0.947**	0.51	0.48	0.413
	NPV	**0.970**	0.82	0.82	0.960
	J	**0.912**	0.03	0.06	0.600
Musculoskeletal Diseases (excluding Spinal Diseases)	Accuracy	**0.924**	0.73	0.72	-
	Sensitivity	**0.922**	0.06	0.09	0.778
	Specificity	0.925	**0.98**	0.97	0.664
	PPV	**0.921**	0.53	0.50	0.215
	NPV	0.927	0.73	0.74	**0.962**
	J	**0.847**	0.04	0.06	0.442
Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb	Accuracy	**0.927**	0.86	0.83	-
	Sensitivity	0.655	0.01	0.12	**0.880**
	Specificity	0.967	**1.00**	0.94	0.816
	PPV	**0.742**	0.43	0.28	0.275
	NPV	0.951	0.86	0.87	**0.988**
	J	0.622	0.01	0.06	**0.696**
Mesothelioma, Asbestosis and Pleural Plaque	Accuracy	**0.993**	0.98	0.98	-
	Sensitivity	**0.934**	0.80	0.81	0.931
	Specificity	**0.996**	0.99	0.99	0.883
	PPV	**0.925**	0.83	0.80	0.425
	NPV	**0.997**	0.99	0.90	0.993
	J	**0.930**	0.79	0.80	0.814

Continued

Ear Disorders (including Hearing Loss)	Accuracy	**0.998**	0.74	0.74	-
	Sensitivity	**0.982**	0.39	0.42	**0.842**
	Specificity	**1.000**	0.89	0.88	0.726
	PPV	**0.997**	0.62	0.61	**0.789**
	NPV	**0.998**	0.77	0.77	0.791
	J	**0.982**	0.28	0.30	0.568

5. Conclusion

This study investigates the application of neural networks for predicting the risk of occupational diseases based on worker and workplace characteristics. An ensemble of one-vs-all classifiers is trained to identify six prevalent disease classes. Model performance is assessed using accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The results show promising performance in most disease classes. For all six disease classes, the specificity values are higher than 0.920 when averaged over 10 runs on the test sets, and for five of the six classes, they exceed 0.967. Regarding sensitivity, the performance is positive (average over 10 runs greater than 0.920) for all classes, except for the classifier for "Carpal Tunnel Syndrome and other Mononeuropathies of Upper Limb", which performs less effectively (accuracy: 0.93, sensitivity: 0.66, specificity: 0.97, PPV: 0.74, NPV: 0.95). Therefore, although the comparison with results from previous studies is encouraging, the classifiers still need improvement on multiple fronts, some of which were mentioned in the previous section. Finally, testing the tool in real-world conditions with occupational physicians could provide valuable insights.

Limitations of the Study

The MalProf System does not cover the entire national territory. Moreover, records are often incomplete or contain inconsistent information. Finally, although the approach of using an overall classifier as an ensemble of specific One-vs-All classifiers allowed for customization of each classifier, this potential has not been fully exploited.

Acknowledgements

This study is based on reports of suspected occupational diseases collected between 2019 and 2023 by the Italian Local Health Authorities participating in the MalProf National Surveillance System. I sincerely thank all the users of the MalProf System who contributed to this study by providing data for the national database.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Mukherjee, C., Gupta, K. and Nallusamy, R. (2012) A Decision Support System for Employee Healthcare. 2012 *Third International Conference on Services in Emerging Markets*, Mysore, 12-15 December 2012, 130-135. https://doi.org/10.1109/icsem.2012.25

[2] Paul, R. and Hoque, A.S.M.L. (2010) Clustering Medical Data to Predict the Likelihood of Diseases. 2010 *Fifth International Conference on Digital Information Management* (*ICDIM*), Thunder Bay, 5-8 July 2010, 44-49. https://doi.org/10.1109/icdim.2010.5664638

[3] Huang, Z.H., Yu, D.H. and Zhao, J.Y. (2000) Application of Neural Networks with Linear and Nonlinear Weights in Occupational Disease Incidence Forecast. *IEEE APCCAS* 2000. 2000 *IEEE Asia-Pacific Conference on Circuits and Systems. Electronic Communication Systems*. (*Cat. No.*00*EX*394), Tianjin, 4-6 December 2000, 383-386. https://doi.org/10.1109/apccas.2000.913515

[4] Yuan, C., Li, G., Peihong, Z. and Li, C. (2010). Artificial Neural Network Modeling of Prevalence of Pneumoconiosis among Workers in Metallurgical Industry—A Case Study. 2010 *International Conference on Intelligent System Design and Engineering Application*, Changsha, 13-14 October 2010, 388-393. https://doi.org/10.1109/isdea.2010.111

[5] Filho, D.V., dos Santos, M.A., Ludermir, T.B. and Silva, M.J. (2002) A Fuzzy Approach to Support a Musculoskeletal Disorders Diagnosis. *VII Brazilian Symposium on Neural Networks*, 2002. *SBRN* 2002. *Proceedings*, Pernambuco, 11-14 November 2002, 154. https://doi.org/10.1109/sbrn.2002.1181461

[6] Martiniano, A., Ferreira, R.P., Sassi, R.J. and Affonso, C. (2012) Application of a Neuro Fuzzy Network in Prediction of Absenteeism at Work. *Iberian Conference on Information Systems and Technologies* (*CISTI*), Madrid, 20-23 June 2012, 1-4.

[7] Liu, H., Tang, Z., Yang, Y., Weng, D., Sun, G., Duan, Z., *et al.* (2009) Identification and Classification of High Risk Groups for Coal Workers' Pneumoconiosis Using an Artificial Neural Network Based on Occupational Histories: A Retrospective Cohort Study. *BMC Public Health*, **9**, Article No. 366. https://doi.org/10.1186/1471-2458-9-366

[8] Srinivas, K., Rao, G.R. and Govardhan, A. (2010) Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques. 2010 5*th International Conference on Computer Science & Education*, Hefei, 24-27 August 2010, 1344-1349. https://doi.org/10.1109/iccse.2010.5593711

[9] Di Noia, A., Montanari, P. and Rizzi, A. (2014) Occupational Diseases Risk Prediction by Cluster Analysis and Genetic Optimization. *Proceedings of the International Conference on Evolutionary Computation Theory and Applications*, Rome, 22-24 October, 68-75. https://doi.org/10.5220/0005077800680075

[10] di Noia, A., Montanari, P. and Rizzi, A. (2015) Occupational Diseases Risk Prediction by Genetic Optimization: Towards a Non-Exclusive Classification Approach. In: Merelo, J.J., Rosa, A., Cadenas, J.M., Dourado, A., Madani, K. and Filipe, J., Eds., *Computational Intelligence*, Springer, 63-77. https://doi.org/10.1007/978-3-319-26393-9_5

[11] Di Noia, A., Martino, A., Montanari, P. and Rizzi, A. (2019) Supervised Machine Learning Techniques and Genetic Optimization for Occupational Diseases Risk Prediction. *Soft Computing*, **24**, 4393-4406. https://doi.org/10.1007/s00500-019-04200-2