



Contents lists available at ScienceDirect

Arabian Journal of Chemistry

journal homepage: www.ksu.edu.sa

Original article

Statistical analysis of the content of metals in blood serum and in the alternative material in head and neck carcinoma



Małgorzata Herman^a, Małgorzata Jakubowska^b, Wojciech Piekoszewski^a, Anna Woźniak^c,
Agnieszka Przybyłowicz^a, Marta Szukalska^c, Ewelina Golusińska-Kardach^d,
Wojciech Gawęcki^e, Agata Czarnywojtek^f, Stanisław Walas^a, Wojciech Golusiński^g,
Marek Ruchała^h, Krzysztof Szyfterⁱ, Alfio Ferlito^j, Ewa Florek^{c,*}

^a Group of Toxicological and Pharmaceutical Analyzes, Department of Analytical Chemistry, Faculty of Chemistry, Jagiellonian University, 2 Gronostajowa St., 30-387 Kraków, Poland

^b Department of Analytical Chemistry, Faculty of Materials and Ceramics, AGH University of Science and Technology, 30 A. Mickiewicza St., 30-059 Kraków, Poland

^c Laboratory of Environmental Research, Department of Toxicology, Poznan University of Medical Sciences, 3 Rokietnicka St., 60-806 Poznań, Poland

^d Department and Clinic of Dental Surgery, Periodontal Diseases and Oral Mucosa, Poznan University of Medical Sciences, 70 Bukowska St., 60-812 Poznań, Poland

^e Department of Otolaryngology and Laryngological Oncology, Poznan University of Medical Sciences, 49 Przybyszewskiego St., 60-355 Poznań, Poland

^f Department of Pharmacology, Poznan University of Medical Sciences, 3 Rokietnicka St., 60-806 Poznań, Poland

^g Department and Clinic of Head and Neck Surgery and Laryngological Oncology, Poznan University of Medical Sciences, 15 Garbary St., 61-866 Poznań, Poland

^h Department and Clinic of Endocrinology, Metabolism and Internal Diseases, Poznan University of Medical Sciences, 49 Przybyszewskiego St., 60-355 Poznań, Poland

ⁱ Institute of Human Genetics, Polish Academy of Sciences, 12 Strzeszyńska St., 60-479 Poznań, Poland

^j Coordinator of the International Head and Neck Scientific Group, 35100 Padua, Italy

ARTICLE INFO

Keywords:

Carcinoma of the head and neck

Hair

Nails

Metals

Statistical analysis

ABSTRACT

The study attempted to demonstrate the differences in the content of essential and toxic metals in the blood serum and in the alternative material (hair and nails) in patients with head and neck cancer compared to healthy people. Selected metals were determined in the test material with the ICP MS technique. The obtained results were subjected to statistical analysis. The Kolmogorov-Smirnov test was used to verify the hypothesis that each tested chemical compound taken from other tissue affects the possibility of developing head and neck cancer or is an indicator of the disease. All variables ($n = 30$) were evaluated to verify significant differences ($p < 0.05$) in each mean between healthy ($n = 55$) and cancer groups ($n = 68$). The data matrix was pre-standardized prior to the analyzes due to large differences in the mean of the compounds tested. The unsupervised learning methods were used and supervised learning methods. On the basis of the statistical analysis, the usefulness of the analysis of alternative material was demonstrated in the potential identification of changes caused in people with cancerous changes at a time when they are not yet visible in the blood serum and can be used as a supplementary test in the diagnosis of head and neck cancer. This is particularly important at early cancer as usually laryngeal cancer is diagnosed in advanced stage.

1. Introduction

The cause of the development of neoplastic diseases is considered in terms of many aspects, including the influence of metals, including heavy metals, and their levels in the body in relation to environmental

factors, smoking, drinking alcohol, diet or the type of professional involvement (Bandeira et al., 2018; Chen et al., 2019; Lv et al., 2022; Mehra and Juneja, 2005; Salcedo-Bellido et al., 2021; Siegel et al., 2014). According to a report prepared by the WHO, only nine metals: iron, zinc, cobalt, nickel, chromium, copper, manganese, molybdenum,

* Corresponding author at: Head of Laboratory of Environmental Research, Department of Toxicology, Poznan University of Medical Sciences, Rokietnicka 3 Street, 60-803 Poznan, Poland.

E-mail addresses: malgorzata.herman@uj.edu.pl (M. Herman), jakubows@agh.edu.pl (M. Jakubowska), wojciech.piekoszewski@uj.edu.pl (W. Piekoszewski), vozniak.anna@gmail.com (A. Woźniak), przybylowicz.agnieszka@gmail.com (A. Przybyłowicz), martan@ump.edu.pl (M. Szukalska), wojgaw@ump.edu.pl (W. Gawęcki), agata.czarnywojtek@ump.edu.pl (A. Czarnywojtek), walas@chemia.uj.edu.pl (S. Walas), wojgaw@ump.edu.pl (W. Golusiński), h.mruchala@ump.edu.pl (M. Ruchała), szyfkris@man.poznan.pl (K. Szyfter), profalfioferlito@gmail.com (A. Ferlito), eflorek@ump.edu.pl (E. Florek).

<https://doi.org/10.1016/j.arabjc.2023.105577>

Received 18 July 2023; Accepted 17 December 2023

Available online 18 December 2023

1878-5352/© 2023 The Authors. Published by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

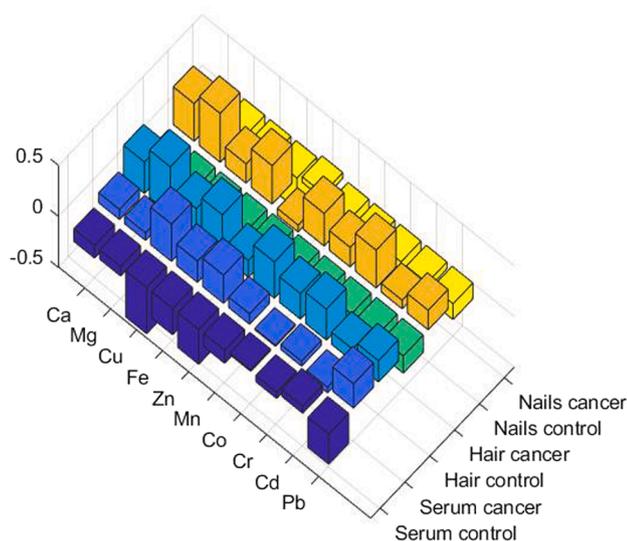


Fig. 1. Mean values of two classes, obtained for data after standardization (the number next to the element symbol has the following meaning: 1 serum, 2 hair, 3 nails.).

selenium are included in the group of essential trace elements (Uthus and Poellot, 1996). Unfortunately, both the deficiency and the excess of trace elements lead to disturbances in the organism's homeostasis ("Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc," 2001; Goldhaber, 2003; Reilly, 2004), and in extreme cases too high concentrations of metals have a toxic effect, causing adverse clinical effects. The biochemical role of trace elements is very complex and multidirectional. They are part of enzymes, hormones, protein carriers, through regulatory proteins to high-energy compounds, they influence the course of physiological mechanisms, regulate organ and systemic functions, and activate the redox system (Goldhaber, 2003). Biological material such as hair and nails has been appreciated in the imaging of various disease entities (Chen et al., 2019; di Ciaula, 2021; G. Nordberg, 2022; Mehra and Juneja, 2005). Unlike blood tests, the composition of which is dynamic and depends on many factors, such as diet and physiological processes aimed at maintaining homeostasis, hair and nails can be used to identify disturbances in this life-giving balance, but can also indicate irregularities that begin. Hence the constant interest of scientists in the search for the dependence of the determined elements, regardless of whether they are essential or toxic, in hair and nails, in relation to popular civilization diseases. Undoubtedly, neoplastic diseases are in one of top place. Indicators of early changes in the body as a result of disease, are looked for and the next step of the research attempts to differentiate between types of illness using subtle differences in metal content, both essential and toxic. The lack of a clear range of reference concentrations, especially for hair, is a certain limitation of the possibility of using an alternative material. Over the course of life, the quality of the hair changes, and therefore the composition of the building material of the hair. Its quality will also depend on sex and growth rate. It must not be forgotten that the place of residence determines environmental factors, and the diet strongly affects the well-being of the body, which has been demonstrated in many studies in the field of science (Barbosa et al., 2005; Thyssen et al., 2011). An example of the antagonistic nature of the interactions between essential elements is the described (Karimi et al., 2012; Goldhaber, 2003).

Among the scientific reports, there is a growing interest in aspects that deal with the problems of oncological laryngology diagnostics, including the influence of metals on pathological changes in the oral cavity, pharynx and larynx sites permanently exposed to environmental and occupational chemical contaminations area. The malignant tumors

arising from mucosal epithelium of these organs are the most common malignancies of the head and neck, and are referred to as a collective name of head and neck squamous cell carcinomas (HNSCCs) (Johnson et al., 2020) HNSCCs are the sixth most common cancers worldwide, with 890,000 new cases and 450,000 deaths in 2018, and their incidence continues to rise and is anticipated to rise by 30% by 2030 (Bray et al., 2018; Ferlay et al., 2019; Johnson et al., 2020). A number of external factors have been demonstrated in the etiopathogenesis of these neoplasms, such as exposure to tobacco-derived carcinogens, excessive alcohol consumption, exposure to environmental pollutants and viral infections: human papilloma virus (HPV) and Epstein-Barr virus (EBV), but the role of the genetic factor is also taken into account (aa, 2000). It is believed that cancers of the oral cavity and larynx are mainly associated with smoking, but in throat cancers the important role of viral infections is emphasized, HPV (primarily HPV-16) in cancers of the oropharynx and EBV in cancers of the nasopharynx (Hennessey et al., 2009; Tsang et al., 2020). The influence of environmental tobacco smoke exposure on the development of HNSCC has also been demonstrated (Mariano et al., 2022; Ramroth et al., 2008) and in developing countries such as India and China exposure to carcinogenic air pollutants is considered a risk factor of these neoplasms (Mishra and Meherotra, 2014; Wong et al., 2014). In addition, in some Asia-Pacific regions, oral cancer has been associated with chewing areca nut products, including 'betel quid', a term used to variety of mixtures containing areca nut (Zhang et al., 2018). In general, the risk of developing HNSCC for men is 2–4-fold higher than for women. The median age of diagnosis for non-virally associated HNSCC is 66 years, whereas for HPV-associated cancer of oropharynx and EBV-associated cancer of nasopharynx is 53 years and 50 years, respectively (Fung et al., 2016; Windon et al., 2018). A very important role in carcinogenesis is played by metals. There is evidence that chronic exposures to heavy metals via tobacco smoking increase the risk of HNSCC (Golasik et al., 2015; Wozniak et al., 2016; Szyfter et al., 2019). Toxic metals have a negative effect on cellular components and enzymes involved in their metabolism, detoxification and repair processes, while the contribution of essential elements can be interpreted in terms of mineral deficiency and the problems associated with it. In this paper, the authors attempted to assess the possibility of differentiating between two groups: patients and volunteers, by comparing the values of the selected elements (Ca, Mg, Cu, Fe, Zn, Mn, Co, Cr, Cd, Pb) in the blood serum and in the alternative material.

2. Experimental

2.1. Patients

The protocol of the study was approved by the Bioethics Committee of the Poznan University of Medical Sciences, Poznan, Poland.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

The analyzed group consisted of two main research groups: study group patients of the Clinic of Head, Neck Surgery and Laryngological Oncology, Clinic of Otolaryngology and Laryngological Oncology Poznan University of Medical Sciences and control group patients of the Clinic of Endocrinology, Metabolism and Internal Diseases Poznan University of Medical Sciences. The control group of 55 patients, including 17 women and 38 men, were patients not diagnosed with cancer and endocrine or metabolic diseases, based on the interview, appropriate laboratory tests and medical examination, they were classified as healthy and not occupationally exposed to heavy metals. The presented results are part of a previously described larger study. This approach used data that included information on both blood serum, hair, and nails. Since not all patients consented to the collection of each type of material or they could not be collected for natural reasons (hair -

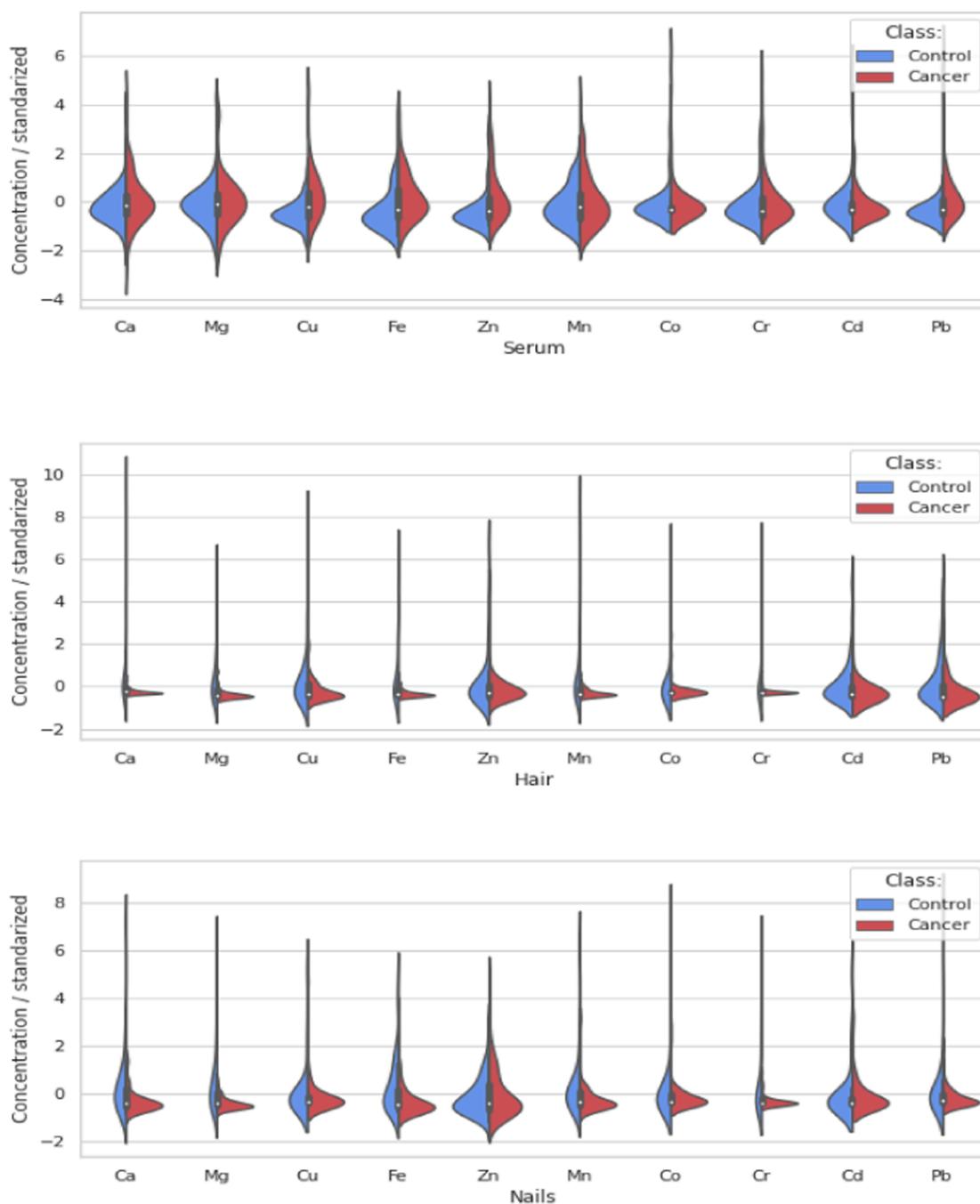


Fig. 2. Violin plots - comparison of the distribution of variables in the control and cancer groups, the concentration of elements: **(top)** in the serum, **(middle)** in the hair, **(bottom)** in the nails.

alopecia or nails - too short), samples were selected for the study in such a way that they contain from each of the selected patients, i.e. hair and/or nails and blood serum. The group of patients consisted of 68 people: 12 women and 56 men. Patients in three age groups were qualified for the study: 20–39, 40–59 and over 60 years old. Both the control group and the patients filled out the same questionnaire with questions about gender and age, as well as lifestyle, amount of alcohol consumed, smoking, diet or medication (these data were not included in this approach). All participants of the study agreed to the examination of the collected material.

2.2. Sample preparation

Certified reference materials for both hair GBW 07601 (GSH-1)

NRCG (China) and serum Seronorm Trace Elements Whole Blood L-2 SERO (Norway) were used. The collected clinical material blood serum stored in a freezer (-80°C) and next thawed at room temperature prior to analysis. Hair from the occipital part of the head and nails (about 0.2 g) were collected and placed in paper envelopes. Then, it was prepared according to the developed method using high-pressure mineralization assisted by microwave radiation. Due to the lack of a reference material for nails on the market, a laboratory nail sample was used for the study. The sample was prepared by grinding material collected from several healthy volunteers. Hair samples were collected from each patient and volunteer from the back of the head (occiput) and nail samples from fingers and toes. The samples were cut with ceramic scissors into 4–5 mm fragments. The nails were stored in plastic vials and the hair in envelopes, both at room temperature. The preparation process included

Table 1
Results of the Kolmogorov–Smirnov test (only variables for which $p < 0.05$).

Element	p-value	Control (mean)	Cancer (mean)	Fold change
Serum				
Cu	$p < 0.001$	859.6	1225.1	0.7
Fe	$p < 0.001$	3100.0	3926.5	0.8
Zn	$p < 0.001$	435.7	762.6	0.6
Cd	$p < 0.005$	7.3	5.8	1.3
Pb	$p < 0.001$	11.6	24.3	0.5
Hair				
Ca	$p < 0.001$	14949.0	1941.0	7.7
Mg	$p < 0.001$	2061.1	301.1	6.8
Cu	$p < 0.01$	87.0	40.4	2.2
Fe	$p < 0.001$	1666.0	257.8	6.5
Mn	$p < 0.001$	15.2	3.3	4.7
Co	$p < 0.001$	1.7	0.3	5.3
Cr	$p < 0.01$	29.2	2.6	11.3
Cd	$p < 0.05$	1.5	1.0	1.5
Pb	$p < 0.005$	5.2	2.8	1.8
Nails				
Ca	$p < 0.001$	2917.9	1591.9	1.8
Mg	$p < 0.001$	974.2	297.9	3.3
Cu	$p < 0.05$	45.0	22.5	2.0
Fe	$p < 0.001$	350.4	151.2	2.3
Mn	$p < 0.001$	5.2	2.1	2.5
Cr	$p < 0.001$	9.6	1.8	5.2
Pb	$p < 0.05$	2.2	1.0	2.1

washing, mixing and grinding the samples in a ball mill. Next, the samples were washed several times with 2 mL of solvent using vortex. The procedure was as follows: water (twice), 1% Triton X-100 (once), water (several times to remove detergent), methanol (once), and finally methanol in an ultrasonic bath for 15 min. After the washing process, the samples were dried at 80°C for several hours (usually overnight). Samples were stored in desiccators. Samples that were stored for more than a month were dried again (30 min, temperature 80°C) before mineralization. The same procedure of the drying procedure was followed in the case of the hair reference material. After turning on the drying process, a 15% improvement in accuracy was observed. Blood samples were collected from the median antecubital vein without anticoagulants and placed in vacuum tubes. The serum was separated by centrifugation at

14,000 rpm for 10 min. The sample serum was stored at -20°C until mineralization. Microwave digestion system (MARS 5X CEM, Matthews, USA) was used to decompose samples. Microwave digestion was performed under optimized conditions. Temperature and pressure were monitored during the heating program. Parameters of the hair and nail mineralization process: temperature 200°C, pressure 38 bar, heating time 15 min, temperature maintenance time 15 min, cooling time 30 min and serum: temperature 242°C, pressure 54 bar, heating time 15 min, temperature maintenance time 15 min, cooling time 30 min. Hair and nail samples of 0.2 g and 0.07–0.1 g, respectively, were taken for analysis, then poured with 7 mL of concentrated nitric acid (V) (which was added to the digestion vessel containing the sample) and left for several hours (usually overnight) before mineralization. For serum, 7 mL

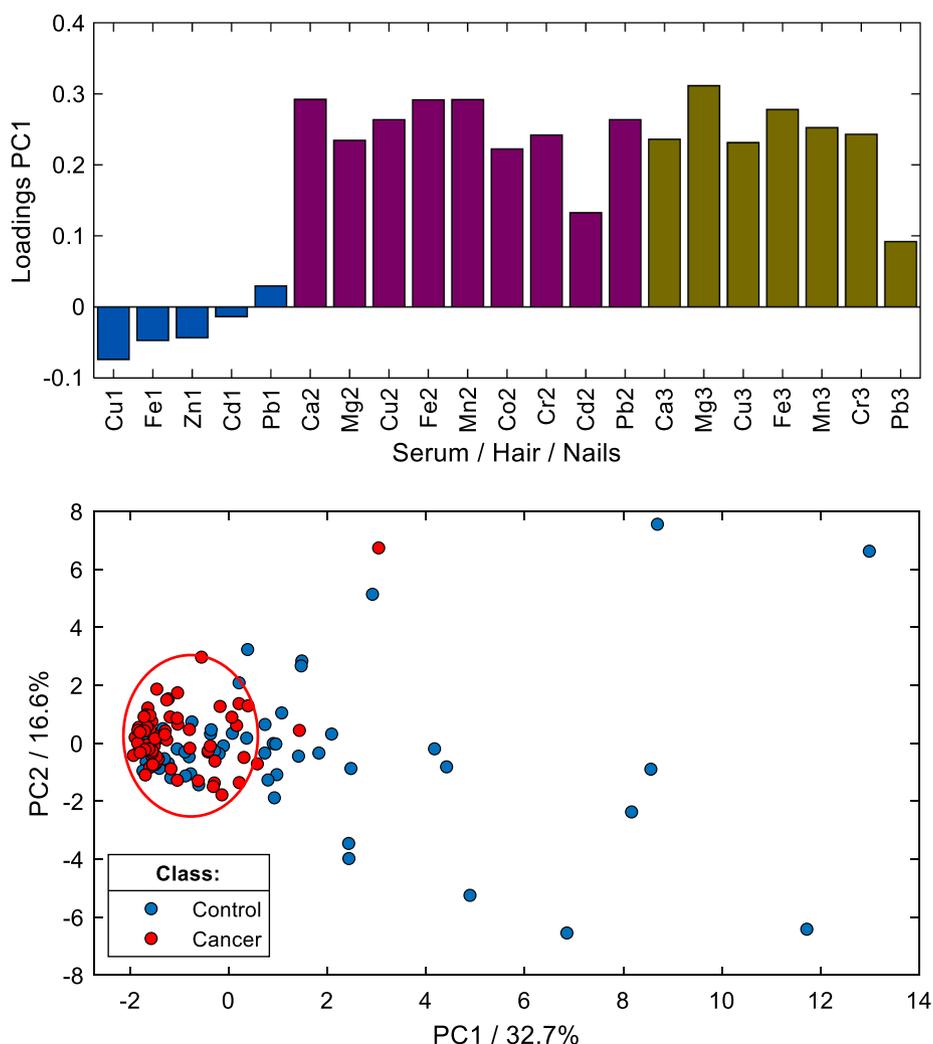


Fig. 3. PCA: (top) projection of objects on the PC1/PC2 plane with the use of 21 class differentiating variables, (bottom) component loadings of PC1. On the X-axis, index 1 is for serum, index 2 for hair, and index 3 for nails.

of concentrated nitric acid (V) was added directly to 0.5 mL of serum samples before digestion. Samples were transferred to 25 mL volumetric flasks and diluted with deionized water. Samples were stored in polyethylene containers in a refrigerator (-4°C).

2.3. Reagents

Only the highest-grade reagents were used. Ultrapure concentrated nitric acid(V) (65%, Suprapur, Merck, Germany) was used for microwave assisted digestion. Methanol, Merck (Germany), 1% Triton X-100 ©, > 99%, Merck (Germany), Water from reverse osmosis process was used. The daily control test of the measuring device was performed using Smart Tune Solution Std ELAN & DRC-e (Perkin Elmer, USA) Atomic Spectroscopy Standard, Multi-element ICP-MS Calibration Std 3, 10 mg L^{-1} , ICP Multi-element Standard Solution VI, 1000 mg L^{-1} , Atomic Spectroscopy Solution, Smart Tune Solution Std ELAN & DRC-e, 10 ppb.

2.4. Apparatus and instruments

During the preparation of the research material and subsequent analysis, the following were used: WG-HLP, Wigo (Poland), shaker with three-dimensional motion, PS-M3D Witko (Poland), Vortex Mixer VX-200 Labnet (USA), ultrasonic cleaner, Sonic 3 Polsonic (Poland), Ball mill, Pulverisette 23, Fritsch GmbH (Germany), 12-position microwave mineralizer equipped with high-pressure XP-1500 vessels, Mars 5X,

CEM, Matthews (USA), multi-station microwave mineralizer with Teflon vessels in a ceramic sheath, Multiwave 3000 Anton Paar (Austria), optical spectrometer with inductively coupled plasma, Optical Emission Spectrometer Optima 2100DV Perkin Elmer (Germany), mass spectrometer with excitation in inductively coupled plasma, ICP MS Mass Spectrometer ELAN DRC-e Axial Field Technology, Perkin Elmer (Germany).

2.5. Methods of chemical analysis

Prior to each analysis the daily control test of the measuring device was performed, next calibration procedure was performed with standard solutions. The reliability of the developed procedures for the determination of selected elements in biological materials was checked during each analysis using CRM. The obtained values were in line with the reference values. All measurements for standards and CRM and analyzed samples were performed in triplicate. The relative errors ranged from 2.56% to max 4.20% in the certified reference materials of hair. The method was characterized by low LOQ, which ranged from 0.009 mg L^{-1} to 0.05 mg L^{-1} for ICP OES and from 0.01 g L^{-1} to 0.69 g L^{-1} for ICP-MS. The intraday precision, expressed as the percent coefficient of variation (CV), was as follows: from 0.51% to 8.02% for hair, from 0.02% to 6.72% for serum and from 0.59% to 6.71% for nails (Golasik et al., 2015).

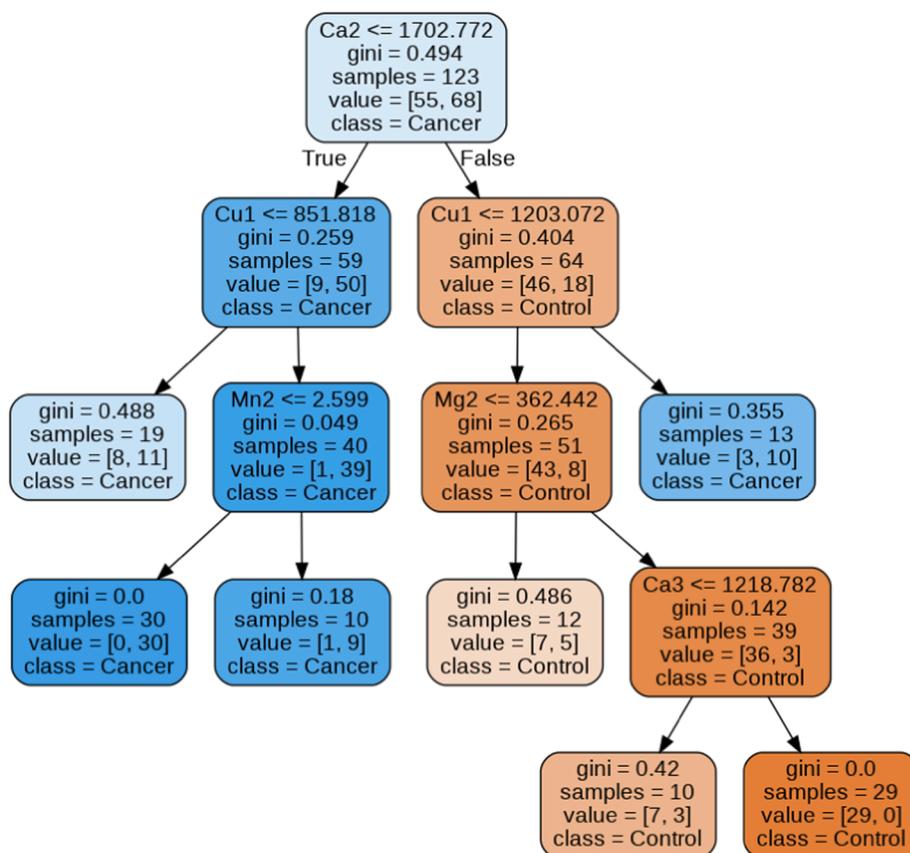


Fig. 4. Decision tree. Shades of blue color in nodes and leaves indicate the cancer group, while shades of orange indicate the control group. Gini shows the impurity of the node (Gini = 0 - clean node, contains objects of one class).

2.6. Statistical analysis

In exploratory data analysis, the Kolmogorov–Smirnov test was applied to verify the hypothesis that each investigated chemical compound (continuous variables) taken from a different tissue affects the possibility of falling head and neck cancer or is an indicator of the disease. Large differences in variation and skewness of the distributions of these compounds justify the choice of the two-sample Kolmogorov–Smirnov test. It was used to verify whether two underlying one-dimensional probability distributions differ and was applied to each variable (Pratt and Gibbons, 1981). All variables ($n = 30$) were evaluated using the Kolmogorov–Smirnov test to verify significant differences ($p < 0.05$) of each mean between the healthy group ($n = 55$) and cancer patients ($n = 68$).

In order to perform comparative research covering all variables and their consistent graphical presentation in charts, the data matrix was initially standardized prior to analyses due to huge differences in means of the investigated compounds. To initially graphically evaluate the differentiation of the concentration distributions of the examined elements in individual tissues and groups, the violin plot was used. It shows the distribution of numerical data across several levels of one (or more) variables such that those distributions can be compared. This can be an effective and attractive way to show multiple distributions of data at the same time. Curve shape corresponds to the approximate frequency of the data points in each region.

2.7. Unsupervised and supervised machine learning

Typically, among machine learning algorithms there are two most important groups, unsupervised learning methods and supervised learning methods (Harrison and Sidey-Gibbons, 2021). Unsupervised

learning methods facilitate the recognition and exploration of the structure of a set of objects, discovery of the primary differentiating factors, grouping of similar cases that form subsets of the analyzed space, and identification of outliers. In this group, one of the most widely used algorithms is Principal Component Analysis (PCA) (Springer, n.d.). This strategy makes it possible to reduce the dimensionality of the data space by transforming the correlated raw variables into new, mutually orthogonal principal components with negligible loss of information. Typically, the first few principal components describe a significant percentage of the information contained in the original data. A significant limitation of the number of variables enables objects to be visualized in a space with a smaller number of dimensions. When a two-dimensional limitation is sufficient, hidden relationships between objects and original variables can be observed on the plane.

Supervised learning is a powerful tool for classifying and processing data using computers. When training models, it uses labeled data, classified by expert knowledge. The effect of training is a model that is then used to predict the unlabeled data that was not exposed during training. The classification approach allows assigning an object characterized by a specific set of measurable features to one of the disjoint groups (classes) of objects or several classes. The basic assumptions of the classification algorithms are the known number of classes and features of objects belonging to each of them, and the classification is qualitative. The problem is to define the rules that allow assigning an object to classes based on the values of explanatory variables that have the highest discriminatory power. The often used classification algorithms with high predictive abilities include the Support Vector Machine (SVM) (Evgeniou and Pontil, 2001), Decision Tree (DT) (“Disease Prediction Model Using Decision Tree Algorithm / 978-620-3-40992-5 / 9,786,203,409,925 / 6203409928,” n.d.; Rokach and Maimon, 2014), Ensemble Methods (Rokach, 2010), together with Random Forest (RF)

Table 2

Accuracy of the classification models using algorithms k-NN, SVM, DT, RF.

Algorithm	Accuracy – cross-validation / %	Accuracy – train set / %	Accuracy – test set / %
<i>Dataset: 30 features</i>			
k-NN	65.8	72.8	71.0
SVM	70.6	77.2	74.2
DT	69.9	84.8	64.5
RF	80.3	88.0	80.7
<i>Feature selection – backward selection</i>			
SVM	76.4	81.5	74.2
RF	80.5	91.3	80.5
<i>Serum – 10 features</i>			
RF	66.6	83.7	67.7
<i>Hair – 10 features</i>			
SVM	72.3	79.3	74.2
<i>Nails – 10 features</i>			
SVM	70.0	75.0	67.7

(Smith, n.d.). For comparison purposes, the Nearest Neighbors (k-NN) algorithm was also applied in this work (Kramer, 2013). An object is classified by a plurality vote of its neighbors, with the object assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). The essence of the SVM method is the construction of an optimal hyperplane, separating data belonging to different classes, with the greatest possible margin of confidence. In the absence of linear separability, data are projected to higher-dimensional spaces, and this strategy consists of selecting an appropriate kernel function, what gives the SVM algorithm a strong advantage over other methods.

Decision trees are an alternative to classical classification methods and do not require a number of applicability conditions. The principle of the algorithm is to create mutually exclusive regions in the space of explanatory variables that contain as many objects from one class as possible. These regions are formed by binary division of sets and subsets of objects according to logical rules, based mainly on one variable at each step. The division continues until homogeneous subsets of objects are obtained and, as a result, information about the division hierarchy is obtained in the form of a tree. An important advantage of models calculated with the use of the DT algorithm is the ease of understanding and visualization of the model and the possibility of interpreting the decision rules. This makes it possible to detect complex interactions between the analyzed variables and objects. The model obtained in this way is not only a tool for the prediction of new cases but also provides a solution to the problem under consideration in a form that is easy for human interpretation.

The DT extension is the random forest algorithm, which creates the model consisting of multiple decision trees calculated in the learning process. This approach is currently included in an intensively developed group of machine learning procedures, i.e., ensemble algorithms.

Ensemble algorithms combine prediction from two or more models to obtain an optimal approach. The ensemble can make better predictions than any single-component model. This is due to the fact that less complex estimators are applied. A random forest is a *meta*-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

The final models and each of the individual classifiers were then applied to the testing dataset to make the final predictions. For each predictive algorithm, the Receiver Operating Characteristic (ROC), Area Under Curve (AUC), accuracy, precision, sensitivity (recall) and confusion matrix was calculated. The ROC and AUC values of the final models were compared (Vujović, 2021).

In this work, the RandomizedSearchCV function (scikit-learn) was used to optimize the hyperparameters of the classification algorithms, which enables randomized search. The parameters of the estimator are optimized by cross-validating search over considered settings. The following hyperparameters of classification algorithms were considered: for k-NN – number of neighbors, for SVM – kernel function, regularization parameter C, kernel coefficient gamma, for DT – the function to measure the quality of a split (Gini or entropy), the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, for RF - parameters similar to DT and the number of trees in the forest.

Validation was carried out using a k-fold cross-validation strategy (k = 6) (Wang and Zheng, 2013) and along with verification in an external test set that consisted of a random 25% of objects (proportional to the number of classes). Cross-validation was performed using the ShuffleSplit function (scikit-learn), which is random permutation cross-validator and yields indices to split data into training and test sets.

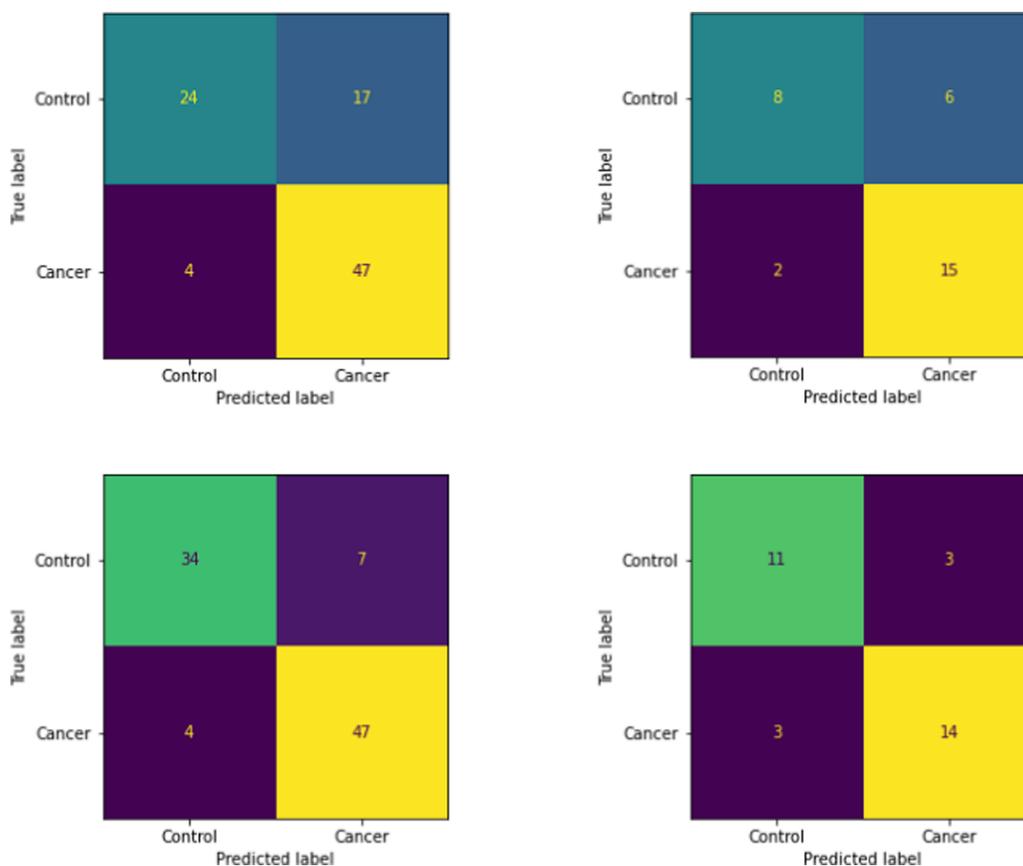


Fig. 5. Confusion matrix: (upper left) SVM, train set, (upper right) SVM, test set, (bottom left) RF, train set (bottom right) SVM, test set.

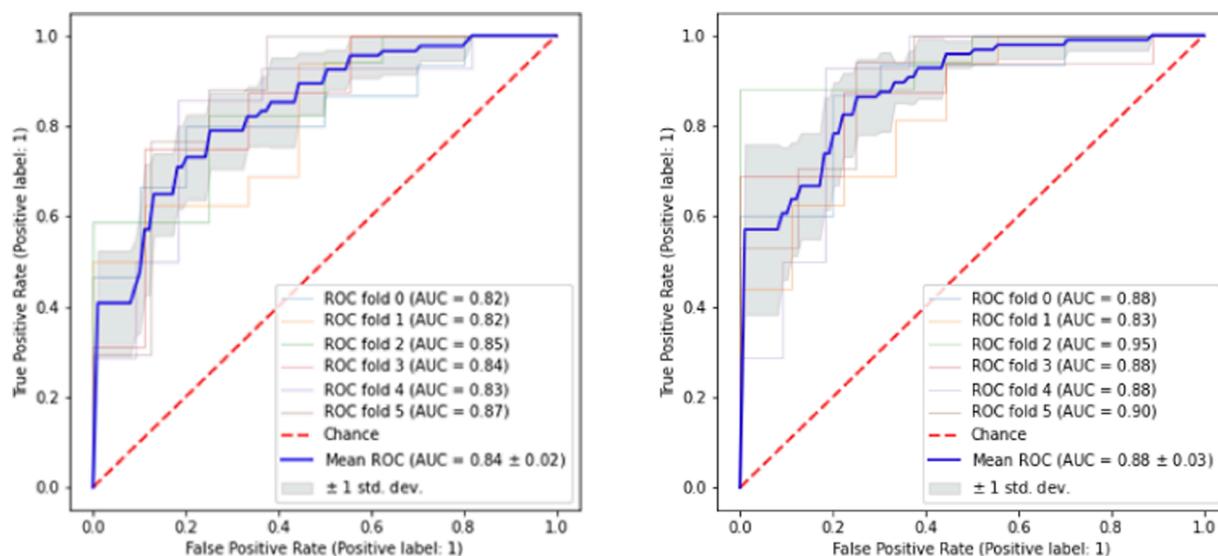


Fig. 6. Results of cross-validation ($k = 6$) of classification models using the ROC curve and the AUC parameter, feature selection using the entire data set (backward selection, 5 features): **left**) SVM, **right**) RF.

2.8. Software

All statistical calculations and PCA were performed in the Matlab R2021a environment (Matworks Inc., US). Implementation of the remaining calculations was performed in Python, using libraries: seaborn v0.12.1, scikit-learn 1.1.2, matplotlib 3.6.1, and an extensive module from the mlxtend 0.21.0 library dedicated to the feature selection, which are accessible under an open-source license.

3. Results and discussion

The data for interpretation were stored in a matrix with 123 rows and 30 columns, where the first dimension represents the number of objects, and the second represents the number of variables (features). The dataset contains the results of the concentrations of ten elements in the serum, hair, and nails of people, of whom 55 belonged to the control group, and the remaining 68 were diagnosed with cancer.

Fig. 1 displays the mean values calculated in two classes after prior standardization of the variables. These values correspond to the concentrations of individual elements, and clear differences are observed between the control and cancer groups as well as between the tested tissues. When considering data relating to the elements in serum, the concentrations of 10 elements in the control group are typically lower than those in the cancer group (with the exception of cadmium and cobalt). Conversely, in hair and nails, the opposite relationship is observed, with the concentrations of the tested analytes being higher in the control group (except for zinc content in nails).

Fig. 2, presented in the form of violin plots, facilitates the analysis of the distributions of all variables separately for two groups in serum, hair, and nails. Each element of the diagram shows the distribution of the tested feature calculated for the control group on the left and the cancer group on the right. In serum, similar concentration distributions of the ten analytes can be observed for the two considered groups. In some cases, the maximum distribution is shifted towards higher values for the cancer group, which confirms previous observations. The results obtained for the concentration of the tested elements in hair allow the conclusion that the cancer group exhibits much smaller variation of the measured values compared to the control group. This is particularly true for Ca, Mg, Fe, Mn, Co, and Cr. Only for Cd and Pb are the distribution characteristics similar. Likewise, the investigated elements in nails show smaller variability for cancer patients, with the exception of zinc, where the distributions are comparable.

The preliminary comparison of classes and features formed the basis for the application of the Kolmogorov-Smirnov test to statistically evaluate differences in distributions. The results are presented in Table 1, taking into account only those variables for which $p < 0.05$. The table lists 21 variables (5 for serum, 9 for hair, 7 for nails) that statistically differentiate the control group from the cancer group.

To confirm the ability of the developed strategy to differentiate objects into two groups - control and cancerous - a principal component analysis was conducted. Fig. 3a illustrates the projection of all objects on the PC1/PC2 plane. The first and second principal components collectively account for 49.3% of the variability (information) present in the dataset. The grouping of cancer objects on the left side of the graph was noticeable, as indicated by an ellipse. These objects exhibited lower values of the first principal component of PC1, implying that this new variable differentiates the two considered classes. To associate this observation with the original variables, PC1 component loadings were examined (Fig. 3b). The loadings for the concentrations of selected elements in hair and nails had much larger weights than for serum. Hence, the first significant conclusion was that the examination of hair and nails enables the differentiation of control and cancerous groups to a greater extent than serum testing.

This was consistent with the earlier observation that the distribution of the variables computed separately in the two considered groups was comparable in the case of the serum and was different for hair and nails. Next, classification models were defined, validated, and optimized in order to develop an efficient and reliable tool for the assessment of new cases based on the tested concentration of selected elements in the blood serum, hair and nails. For the four selected algorithms, the optimal hyperparameters were: for k-NN - number of neighbors = 6, for SVM - radial basis kernel function, $C = 40$, for DT - the function to measure the quality of a split - Gini, the maximum depth of the tree = 5, the minimum number of samples required to split an internal node = 2, the minimum number of samples required to be in a leaf node = 10, for RF - parameters the same as for DT and the number of trees in the forest = 100.

The classification models were evaluated using the ROC curve and AUC coefficients. The obtained AUC values represent the area under the ROC curve, which reflected the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), i.e., it directly presented the assessment of the predictive capacity of the models for cancer patients. The highest AUC values were obtained for the model calculated using

SVM (0.84 ± 0.02) and RF (0.88 ± 0.04). Low standard deviation values were noteworthy, which proved the ability of the models to fit the data.

Fig. 4 shows a decision tree, which was a classification model and at the same time allowed for inference about the significance of individual features. In the root, a binary division was realized on the basis of Ca in the hair, which suggested that the concentration of this element is important in the diagnostic procedure.

Table 2 presents the evaluation parameters of the classification models defined using a cross-validation strategy, for the training set (75% of cases) and validated on the external set (25%) of the objects. The first four rows refer to the calculation variant when all features (30) were included in the definition of the classification model.

Additionally, for the two best models, defined with the use of SVM and RF algorithms, which had the ability to correctly predict new objects at the level of 74.2% and 80.7%, a confusion matrix was calculated and presented (Fig. 5), separately for the training set and the test set. They allowed the assessment of how the algorithms worked for individual groups i.e. control and cancer. For these models, typical quality evaluation parameters were also obtained, i.e. precision, recall and f1-score. The desired compliance of the parameters for both groups in the case of the model using the RF formalism was significant. In this case, the precision, recall and f1-score for the control group was 0.79, and for the cancer group the values of these scores were equal 0.82.

An iterative feature selection algorithm called backward selection was applied next. The procedure selected 5 significant features that allowed for the definition of models with similar quality as the entire set of features. In the case of the SVM algorithm, these features were copper (Cu1) and iron (Fe1) concentration in serum, calcium (Ca2) and magnesium (Mg2) concentration in the hair, and the concentration of magnesium (Mg3) in nails. On the other hand, when the predictive model was defined with RF, the important features were calcium (Ca1), copper (Cu1), and zinc (Zn1) concentration in serum, calcium (Ca2) in the hair, and the concentration of iron (Fe3) in the nails. When comparing the mean AUC values, they were identical to those for 30 features, but the variability of the AUC expressed by the standard deviation was smaller. ROC charts and cross-validation results are shown in Fig. 6. After selecting the most important features, the model defined with the use of RF algorithms achieved high predictive ability for new objects, with 80.5% correct classifications (Table 2). These observations and conclusions suggest that testing a smaller number of characteristics of the material taken from patients is sufficient. In this case, the level of probability of a correct diagnosis is at the same level as for the full list of features considered in this work, which is 30. Contents of a short list of material characteristics depends on which classification algorithm is used in modeling the responses.

The calculations were continued by defining the models, using separately the data for serum, hair, and nails. Overall, it can be concluded that by examining the concentrations of 10 elements in the hair, high-quality predictive models were obtained for the various algorithms considered. When only serum features were included in the calculations, the k-NN and SVM models were characterized by low quality and low predictive abilities. For nails, their characteristics can be used to define a high-quality predictive model using the SVM algorithm. A detailed assessment of the models' accuracy is given in Table 2.

Similar statistical studies illustrating the usefulness of the use of alternative material, apart from blood samples, have been conducted for many years, but relationships between groups of patients of various disease entities and control groups in relation to various analytes are still being sought. For example, Qayyum and Shah (Qayyum and Shah, 2017) showed differences in the study material between essential and toxic trace metals (Cd, Cr, Cu, Fe, Mn, Ni, Pb and Zn) in patients with oral cancer and in the control group. Significantly higher average concentrations of Cd, Ni and Pb ($p < 0.05$) as well as Cu, Fe and Zn were found in the blood, hair and nails of patients compared to the control group. They used a multivariate cluster analysis of metal levels and showed that the study illustrates significant differences in metal levels in oral cancer

patients compared to controls.

4. Conclusions

Serum samples and an alternative material such as hair and nails were included in the considerations. It has been shown to be useful in the potential identification of neoplastic lesions of the head and neck at a time when they are not yet visible in the blood serum.

The usefulness of the tests used as an aid in clinical diagnosis has been demonstrated. For example, the Decision tree test used differentiates the pool of people included in the study into groups of healthy and potentially sick people, indicating the importance of analyzing the material in terms of the determination of elements such as: Ca, Cu, Mn, Mg. These elements also appear as significant in the Kolmogorov–Smirnov test used.

Based on the calculations performed, it can be concluded that the entire data set as well as its subsets selected in a targeted manner allow for automatic inference and prediction of belonging to the group of healthy or cancerous. Statistical analysis indicated that 21 examined features as significantly differentiating the two classes, while the multivariate approach indicated even a smaller number of characteristics as significant. PCA and supervised learning allow the conclusion that the concentrations of elements in the hair and nails are primarily useful in research. It is obvious that early detection of tumor is useful for efficient therapy and contributes to beneficial prognosis. Noninvasive way of material gaining is another worth to mention advantage of the presented attempt to diagnosis. The model defined with the use of the random forest algorithm, computed using the entire set, achieved the highest predictive abilities, i.e. accuracy 80.7%.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The study was approved by the Bioethics Committee of the Poznan University of Medical Sciences. The study was approved by the Bioethics Committee of the Poznan University of Medical Sciences, Poznan, Poland. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Acknowledgement

The financial support under research grant number UMO-2011/03/N/NZ7/06266 by The National Science Centre and Poznan University of Medical Sciences grant number 004648 is kindly acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.arabjc.2023.105577>.

References

- Bandeira, C.M., de Almeida, A.Á., Carta, C.F.L., Almeida, A.A., de Figueiredo, F.A.T., Sandrim, V.C., Gonçalves, A.J., Almeida, J.D., 2018. Tobacco influence in heavy metals levels in head and neck cancer cases. *Environ. Sci. Pollut. Res.* 25, 27650–27656. <https://doi.org/10.1007/s11356-018-2668-9>.
- Barbosa, F., Tanus-Santos, J.E., Gerlach, R.F., Parsons, P.J., 2005. A critical review of biomarkers used for monitoring human exposure to lead: Advantages, limitations, and future needs. *Environ. Health Perspect.* <https://doi.org/10.1289/ehp.7917>.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. <https://doi.org/10.3322/CAAC.21492>.
- Chen, Q.Y., DesMarais, T., Costa, M., 2019. Metals and Mechanisms of Carcinogenesis. *Annu. Rev. Pharmacol. Toxicol.* 59, 537–554. <https://doi.org/10.1146/ANNUREV-PHARMTOX-010818-021031>.
- di Ciaula, A., 2021. Bioaccumulation of Toxic Metals in Children Exposed to Urban Pollution and to Cement Plant Emissions 13, 681–695. <https://doi.org/10.1007/s12403-021-00412-w>.
- Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc, 2001. . . Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc. Doi: 10.17226/10026.
- Disease Prediction Model Using Decision Tree Algorithm / 978-620-3-40992-5 / 9786203409925 / 6203409928 [WWW Document], n.d. URL <https://www.lap-publishing.com/catalog/details/store/gb/book/978-620-3-40992-5/disease-prediction-model-using-decision-tree-algorithm> (accessed 12.2.22).
- Evgeniou, T., Pontil, M., 2001. Support vector machines: Theory and applications. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2049 LNAI, 249–257. Doi: 10.1007/3-540-44673-7_12/COVER.
- Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.M., Piñeros, M., Znaor, A., Bray, F., 2019. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* 144, 1941–1953. <https://doi.org/10.1002/IJC.31937>.
- Fung, S.Y.H., Lam, J.W.K., Chan, K.C.A., 2016. Clinical utility of circulating Epstein-Barr virus DNA analysis for the management of nasopharyngeal carcinoma. *Chin. Clin. Oncol.* 5 <https://doi.org/10.21037/CCO.2016.03.07>.
- Golasik, M., Przybyłowicz, A., Woźniak, A., Herman, M., Gawecki, W., Golusiński, W., Walas, S., Krejpcio, Z., Szyfter, K., Florek, E., Piekoszowski, W., 2015. Essential metals profile of the hair and nails of patients with laryngeal cancer. *J. Trace Elem. Med. Biol.* 31, 67–73. <https://doi.org/10.1016/j.jtemb.2015.03.001>.
- Goldhaber, S.B., 2003. Trace element risk assessment: essentiality vs. toxicity. *Regul. Toxicol. Pharm.* 38, 232–242. [https://doi.org/10.1016/S0273-2300\(02\)00020-X](https://doi.org/10.1016/S0273-2300(02)00020-X).
- Harrison, C.J., Sidey-Gibbons, C.J., 2021. A systematic review of the quality of conduct and reporting of survival analyses of tuberculosis outcomes in Africa. *BMC Med. Res. Method.* 21, 158. <https://doi.org/10.1186/s12874-021-01347-1>.
- Hennessey, P.T., Westra, W.H., Califano, J.A., 2009. Human papillomavirus and head and neck squamous cell carcinoma: recent evidence and clinical implications. *J. Dent. Res.* 88, 300–306. <https://doi.org/10.1177/0022034509333371>.
- Johnson, D.E., Burtness, B., Leemans, C.R., Lui, V.W.Y., Bauman, J.E., Grandis, J.R., 2020. Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Primers* 6. <https://doi.org/10.1038/S41572-020-00224-3>.
- Karimi, G., Shahar, S., Homayouni, N., Rajikan, R., Abu Bakar, N.F., Othman, M.S., 2012. Association between trace element and heavy metal levels in hair and nail with prostate cancer. *Asian Pac. J. Cancer Prev.* 13, 4249–4253. <https://doi.org/10.7314/apjcp.2012.13.9.4249>.
- Kramer, O., 2013. K-Nearest Neighbors 13–23. Doi: 10.1007/978-3-642-38652-7_2.
- Lv, L., Liu, B., Yu, Y., Dong, W., Gao, L., He, Y., 2022. Heavy metals in paired samples of hair and nails in China: occurrence, sources and health risk assessment. *Environ. Geochem. Health.* <https://doi.org/10.1007/S10653-022-01400-Y>.
- Mariano, L.C., Warnakulasuriya, S., Straif, K., Monteiro, L., 2022. Secondhand smoke exposure and oral cancer risk: a systematic review and meta-analysis. *Tob. Control* 31, 597–607. <https://doi.org/10.1136/TOBACCOCONTROL-2020-056393>.
- Mehra, R., Juneja, M., 2005. Fingernails as biological indices of metal exposure. *J. Biosci. Mishra, A., Meherotra, R., 2014. Head and neck cancer: global burden and regional trends in India. Asian Pac. J. Cancer Prev.* 15, 537–550. <https://doi.org/10.7314/APJCP.2014.15.2.537>.
- G. Nordberg, M.C., 2022. HANDBOOK ON THE TOXICOLOGY OF METALS general considerations.
- Pratt, J.W., Gibbons, J.D., 1981. Kolmogorov-Smirnov Two-Sample Tests 318–344. Doi: 10.1007/978-1-4612-5931-2_7.
- Qayyum, M.A., Shah, M.H., 2017. Study of trace metal imbalances in the blood, scalp hair and nails of oral cancer patients from Pakistan. *Sci. Total Environ.* 593–594, 191–201. <https://doi.org/10.1016/J.SCITOTENV.2017.03.169>.
- Ramroth, H., Dietz, A., Becher, H., 2008. Environmental tobacco smoke and laryngeal cancer: results from a population-based case-control study. *Eur. Arch. Otorhinolaryngol.* 265, 1367–1371. <https://doi.org/10.1007/S00405-008-0651-7>.
- Reilly, C., 2004. The nutritional trace metals. The nutritional trace metals.
- Rokach, L., 2010. Ensemble-based classifiers. *Artif. Intell. Rev.* 33, 1–39. <https://doi.org/10.1007/S10462-009-9124-7>.
- Rokach, L., Maimon, O., 2014. Data Mining with Decision Trees: Theory and Applications, 2nd Edition. *Data Mining with Decision Trees: Theory and Applications*, 2nd Edition 81, 1–305. Doi: 10.1142/9097/SUPPL_FILE/9097_CHAP01.PDF.
- Salcedo-Bellido, I., Gutiérrez-González, E., García-Esquinas, E., Fernández de Larrea-Baz, N., Navas-Acien, A., Téllez-Plaza, M., Pastor-Barriso, R., Lope, V., Gómez-Ariza, J.L., García-Barrera, T., Pollán, M., Jiménez Moleón, J.J., Pérez-Gómez, B., 2021. Toxic metals in toenails as biomarkers of exposure: A review. *Environ. Res.* 197 <https://doi.org/10.1016/j.envres.2021.111028>.
- Siegel, R., Ma, J., Zou, Z., Jemal, A., 2014. Cancer Statistics, 2014. *CA Cancer J. Clin.* 64, 9–29. <https://doi.org/10.3322/caac.21208>.
- Smith, C., n.d. Decision trees and random forests : a visual introduction for beginners 151.
- Springer, I.T.J., n.d. Principal Component Analysis, Second Edition.
- Szyfter, K., Napierala, M., Florek, E., Braakhuys, B.J.M., Takes, R.P., Rodrigo, J.P., Rinaldo, A., Silver, C.E., Ferlito, A., 2019. Molecular and health effects in the upper respiratory tract associated with tobacco smoking other than cigarettes. *Int. J. Cancer* 144 (11), 2635–2643. <https://doi.org/10.1002/ijc.31846>.

- Thyssen, J.P., Roeske-Nielsen, A., Johansen, J.D., 2011. Contact allergy and human biomonitoring – an overview with a focus on metals. *Contact Dermatitis* 65, 125–137. <https://doi.org/10.1111/J.1600-0536.2011.01933.X>.
- Tsang, C.M., Lui, V.W.Y., Bruce, J.P., Pugh, T.J., Lo, K.W., 2020. Translational genomics of nasopharyngeal cancer. *Semin. Cancer Biol.* 61, 84–100. <https://doi.org/10.1016/J.SEMCANCER.2019.09.006>.
- Uthus, E.O., Poellot, R.A., 1996. Dietary folate affects the response of rats to nickel deprivation. *Biol. Trace Elem. Res.* 52, 23–35. <https://doi.org/10.1007/BF02784087>.
- Vujović, Ž.D., n.d. Classification Model Evaluation Metrics. *IJACSA) International Journal of Advanced Computer Science and Applications* 12, 2021.
- Wang, H., Zheng, H., 2013. Model Validation, Machine Learning. *Encyclopedia Syst. Biol.* 1406–1407 https://doi.org/10.1007/978-1-4419-9863-7_233.
- Windon, M.J., D'Souza, G., Rettig, E.M., Westra, W.H., van Zante, A., Wang, S.J., Ryan, W.R., Mydlarz, W.K., Ha, P.K., Miles, B.A., Koch, W., Gourin, C., Eisele, D.W., Fakhry, C., 2018. Increasing prevalence of human papillomavirus-positive oropharyngeal cancers among older adults. *Cancer* 124, 2993–2999. <https://doi.org/10.1002/CNCR.31385>.
- Wong, I.C.K., Ng, Y.K., Lui, V.W.Y., 2014. Cancers of the lung, head and neck on the rise: perspectives on the genotoxicity of air pollution. *Chin. J. Cancer* 33, 476–480. <https://doi.org/10.5732/CJC.014.10093>.
- Wozniak, A., Napierala, M., Golasik, M., Herman, M., Walas, S., Piekoszewski, W., Szyfter, W., Szyfter, K., Golusinski, W., Baralkiewicz, D., Florek, E., 2016. Metal concentrations in hair of patients with various head and neck cancers as a diagnostic aid. *Biomaterials: an Int. J. Role of Met. Ions in Biol., Biochem. Med.* 29 (1), 81–93. <https://doi.org/10.1007/s10534-015-9899-8>.
- Zhang, L.W., Li, J., Cong, X., Hu, X.S., Li, D., Wu, L.L., Hua, H., Yu, G.Y., Kerr, A.R., 2018. Incidence and mortality trends in oral and oropharyngeal cancers in China, 2005–2013. *Cancer Epidemiol.* 57, 120–126. <https://doi.org/10.1016/J.CANEP.2018.10.014>.