

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
INSTITUTE OF COMPUTER SCIENCE
DEPARTMENT OF COMPUTATIONAL AND DATA MODELING
AND
CYBERSECURITY LABORATORY

Final Master Thesis

Research of Computational Methods for Digital Phenotyping for Cybersecurity Risk Assessment

Done by:
Aistė Jurevičienė signature

Supervisor:
Assoc. prof. dr. Linas Bukauskas

Vilnius
2021

Contents

Abstract **3**

Santrauka **4**

Introduction **5**

1 Related Works **8**
1.1 Methods and systems of face parts recognition 8
1.2 Cyberattacks which are directly related to the human being 12
1.2.1 Cyberattacks 13
1.2.2 Human traits that can be used for cyberattacks 15
1.2.3 Risk assessment 17

2 Model **22**
2.1 Ontology-based model 22
2.2 Meta-schema of traits network 24

3 Implementation **27**
3.1 Architecture 27
3.2 Algorithms 29
3.3 Database 31

4 Experiments **32**
4.1 Kinship identification 32
4.2 For risk assessment extracted traits from collected data 38
4.2.1 Social behaviour 39
4.2.2 Traits from photos 46
4.2.3 Members of family recognition 48
4.2.4 Traits correlation 49

Results and Conclusions **57**

Future Works **59**

References **60**

Appendices **67**

A Semantic network **67**

Abstract

People are the most vulnerable part of cyberattacks, so companies and individuals should consider threats arising from human traits. Many characteristics can describe a person; therefore it is essential to identify which of them can be identified by computer technology and might impact security. In this work, we form the semantic network that structures the automatically recognized traits. Based on the semantic network, the research conducted in this work showed that family members are characterized not only by physical but also by character traits. The identification of family members with a trained network is also an essential part of assessing cybersecurity risks. Some trends between property correlations were also observed. People with a higher extrovert score, for example, are more agreeable and less neurotic. Lithuanians, who participated in this research, are extroverted, open, quite conscientious, and agreeable; they make decisions depending on the situation and rather not take many risks. Similar trends are observed between both genders in different age groups. Although generalizations are possible for groups of people, individualized cyber risk assessment can highlight characteristics that are not specific to a group and draw a person's attention to its weaknesses.

Keywords: semantic network model, digital phenotyping, risk assessment, kinship recognition.

Scientific publishing:

- A. Jurevičienė, A. Brilingaitė, and L. Bukauskas, *Digital Human in Cybersecurity Risk Assessment*, accepted to HCI 2021: Human-Computer Interaction, publishing Springer Computer Science Proceedings, LNCS series, Washington DC, USA, July 24–29, 16 pages, 2021.

Santrauka

Analizės algoritmų tyrimas skaitmeniniam fenotipavimui kibernetinio saugumo rizikos vertinimo kontekste

Kibernetinių atakų skaičius auga, nors technologijos tobulėja ir atrodo, jog lieka vis mažiau saugumo spragų, tačiau dėl nepakankamo dėmesio skiriamo švietimui, žmogus išlieka silpnoji grandis. Tobulėjant technologijoms vis daugiau informacijos apie būsimą atakos auką galima nustatyti iš socialiniuose tinkluose pateikiamos informacijos: nuotraukų, vaizdinių įrašų, minčių išreikštų kompiuterinėje erdvėje. Vienos savybės turi daugiau įtakos negu kitos. Šiame darbe pateikiama analizė tam tikrų bruožų kaip amžiaus, lyties, Didžiojo penketo savybių įtaka kibernetiniai higienai ir kitiems veiksniams susijusiems su kibernetiniu saugumu. Identifikuoti žmogaus silpnąsias vietas yra pirminis aspektas, kurį turėtų įvertinti įmonės ar individualūs asmenys, kad būtų galima suteikti tikslingą informaciją padėsiančią išvengti ar bent jau sumažinti kibernetinės atakos tikimybę.

Šiuo magistriniu darbu siekiama sukurti metodiką ir skaičiavimo modelius žmogaus fenotipui nustatyti pagal skaitmeninius ar stebėjimo duomenis, siekiant įvertinti kibernetinio saugumo riziką. Tikslui pasiekti iškelti keturi uždaviniai:

1. Ištirti žmogaus bruožų tyrimų algoritmus ir technologinius sprendimus.
2. Apibrėžti ir sukurti savybių taksonomiją galimam kibernetinio saugumo atakos nustatymui.
3. Išplėtoti rizikos vertinimo modelį paremtą ontologija.
4. Sumodeliuoti prototipinę sistemą, kuri nustatytų galimą kibernetinės rizikos balą.

Remiantis literatūros šaltiniais sudarytas žmogaus bruožų semantinis tinklas, kai kurioms savybėms suteikti kibernetinės rizikos vertinimo įverčiai, sumodeliuotas ontologija paremtas modelis. Gautų struktūrizuotų duomenų pagrindu sudaryta prototipinė rizikos vertinimo sistema. Atliktas tyrimas su 22 šeimų nariais, įdentifikuojant individualius bruožus, giminystės ryšį iš nuotraukų, nustatant charakterio ypatybes pagal anketos klausimus. Kraujo ryšiui atpažinti apmokytas neuroninis tinklas iš „Deepface“ projekto, tačiau atpažinimo tikslumas yra tobulintinas. Iš gautų tyrimo rezultatų pastebimos tam tikros bruožų koreliacijos. Taip pat, darytina išvada, kad lietuviai, dalyvavę tyrime, yra ekstravertiški, atviri patirčiai, gana nuolaidūs ir stropūs. Linkę priimti sprendimus priklausomai nuo situacijos ir nelinkę rizikuoti. Tačiau didžiajai daliai trūksta žinių apie kibernetinį saugumą. Tai tik rodo kibernetinio rizikos vertinimo svarbą, siekiant sudaryti sąlygas supažindinti žmones su jų pažeidžiamumais.

Raktažodžiai: semantinio tinklo modelis, skaitmeninis fenotipavimas, rizikos vertinimas, kraujo ryšių atpažinimas.

Mokslinis publikavimas:

- A. Jurevičienė, A. Brilingaitė, and L. Bukauskas, *Digital Human in Cybersecurity Risk Assessment*, priimta į HCI 2021: Human-Computer Interaction, leidėjas Springer Computer Science Proceedings, LNCS series, Washington DC, USA, July 24–29, 16 psl., 2021.

Introduction

In cyberattacks, the most vulnerable part is human. Knowing a person's weaknesses makes it easy to gain access to information for the attacker; therefore, social engineering, spear-phishing is very effective. After receiving data from one employee, an attacker can usually access other data in the company's system. That situation arose in 2012 [12] when one employee of Dropbox used the same password to login to Dropbox and LinkedIn. The phisher used a password and gained access to the company's network and database with users' emails. By linking human phenotypic traits to genes, can be made more accurate assumptions about human characteristics, behavior in certain situations. Machine learning from face picture helps to identify whether a person is suffering from a syndrome or not. Based on measurements of facial features, human behavior in the environment, and other characteristics, we can predict whether people tend to take risks, how good metabolize of xenobiotics or have some psychological problems.

People who spend less time on computers, women, seniors exhibit inferior cyber hygiene. Compared to 2013 and 2018, the number of security breaches increased by 67 %, which causes enormous financial damage to companies and individuals. As cyberattacks against medical institutions increase [37], from 2019, 70 % of US states will require nurses to undergo security training to maintain or renew their licenses. Nurses make up the vast majority of medical staff, so they must be trained, that they timely notice the beginning of the cybercrime, or do not become the cyberattacks victim. Not only patient information leaks should be prevented, but events that may disrupt the treatment process or even lead to death should be avoided. Educating people, raising awareness of personalized hazards can help prevent cyberattacks, respond appropriately.

On social networks people reveal a great deal of personal information about themselves and their family members. The attacker analyzes not only the human being but also his friends, the environment. The article "Open-source intelligence for risk assessment" [29] describes risk assessment based on the environment, adding each person close to you with an additional risk score. In critical situations, most people tend to disclose sensitive information to their loved ones. But in personalized attacks, it is essential how exactly the right person will behave in a given situation. Neuroticism has been found to have a positive influence on the success of phishing [24]. If we look at neuroticism at the genetic level [52], it correlates with depression, schizophrenia, ever-smoker. And schizophrenia, like bipolar syndrome, affects facial proportions [30]. Thus, based on the data obtained from digital data such as facial symmetry or observations of frequent smoking, we can derive a broader spectrum of human features by reverse phenotyping.

Human is a part of which, violated, can destroy not only him but also companies, military missions. Many companies do human screening based on human features, psychological tests. Still, there are areas where hiring the right employee is particularly important, as it can cost the leaks of sensitive information or even the lives. A system that based on data available to the general public recognizes data beyond what is evident to everyone would be an excellent tool for identifying a person's vulnerabilities. Certainly, the system would not replace the costly genetic tests, which can indicate a person's possibility to suffer from certain diseases, their relationship with relatives, their ability to metabolize certain substances. However, the results of genetic testing may not be able to tell everything whats will be happening in human life. For example, through pictures with machine learning is identifying—gender—which reflected in genetic information that can not be denied. Nevertheless, whether a person will have a particular disease—genetic information indicates just possibility because there are too many factors such an environment, nutrition. However, it is enough to know the traits that are occurring now or in the near future.

Data that can be collected at a given moment, through analysis of photos, comments on social networks, information from smartwatches or phones about gait, activity, voice, is sufficient to determine human vulnerability. The information obtained can be used for the selection of people in companies or for the security of one's own cyber.

The work aims to develop methodology and computational models for the digital phenotyping of a human from digital or observation data to assess the risk in cybersecurity relating to a cyber kill chain. To achieve the aim we establish goals:

1. Investigate research and technology solutions of human traits recognition.
2. Define and develop a taxonomy of traits for possible cybersecurity attack determination.
3. Develop risk assessment model based on ontology.
4. Model the prototype software as an expert system that evaluates potential cybersecurity risk score.

After analyzing the literature, a semantic network defining the taxonomy of traits was formed. From photos, machine learning methods can be used to determine age, gender, ethnicity, specific skin marks, body complexity, or even sexual orientation. From comments on social networks—a person's psychological state: depression, eating disorders or suicidal ideation. From the data received from smart watches—activity, gender, possible signs of illness or trauma, habits. The semantic network in this work is generalized to the ontology-based model. Graph schema consists of many property paths, which described human risk.

Based on the articles read, a table of scores was compiled according to human phenotypic traits. Characteristics such as female gender, young age, especially between the ages of 18–25, neuroticism, openness, avoidance decision-making style, ethical and healthy/safety risk-taking preference correlate with increased cybersecurity risk. People with these characteristics are susceptible to phishing attacks, have poor cyber hygiene: use of easy-to-crack passwords, lack of computer equipment, telephone securement, non-timely periodic updates, unsafe use of websites. Agreeableness, extraversion, conscientiousness, rational-decision making style, financial and social risk-taking preferences negatively correlate with cybersecurity risk. Photos and answers to the questionnaire are used in the system to identify the mentioned features for risk score calculation. The neural network from the pictures also recognize if people have a kinship, each family member increases family cyber risk by making it easier to access the victim for an attacker. Identification of family members can reveal not only obvious things but also genetically inherited traits. Therefore, say an attacker has little information about the victim, but knowing human relatives can make general assumptions about the victim. A system that analyzes cybersecurity risks by examining human traits reveals features that increase risk and is a great tool to draw a person's attention to weaknesses that can be exploited.

The paper consists of a review of the literature used in the Scientific Research Project: in Subsection 1.1 methods of face detection and recognition are introduced, in Subsection 1.2 cyber attack threat statistics are provided, the main human-related attacks, and how to assess risk based on data collected with machine learning algorithms about humans. And new parts of Final Master Thesis Related Works Section is Subsubsection 1.2.2, where the effects of human characteristics on cyber hygiene and on the success of cyber attacks is discussed. In Model Section 2 we describe ontology-based model, reviewed what information about a person can be collected from social networks, publicly available photos, and generalized it all into meta-schema. The Section 3 details

the implementation of a web application prototype for cyber risk self-assessment. Experimental section consist of subsections: in Subsection 4.1 several trained neural networks to recognize kinship and their efficiency metrics are described. The results discussed in the Subsection 4.2 were obtained by analyzing the data of the subjects in the study.

1 Related Works

1.1 Methods and systems of face parts recognition

In the early 1960 computer scientists started programming computers to recognize human faces. Woodrow W. Bledsoe was one of the first who published his work [10]. Through the years face recognition technologies have improved greatly. Nowadays these technologies are used in the big corporation as "Toyota" (sleeping driver detector for safety reasons) [79] or "Facebook" (face recognition in the uploaded photos). More face recognition application areas see Table 1.

Table 1. Applications of face recognition [10]

Areas	Applications
Information Security	Access security (OS, data bases) Data privacy (e.g. medical records) User authentication (trading, on line banking)
Access management	Secure access authentication (restricted facilities) Permission based systems Access log or audit trails
Biometrics	Person identification (national IDs, Passports, voter registrations, driver licenses) Automated identity verification (border controls)
Law Enforcement	Video surveillance Suspect identification Suspect tracking (investigation) Simulated aging) Forensic Reconstruction of faces from remains
Personal security	Home video surveillance systems Expression interpretation (driver monitoring system)
Entertainment-Leisure	Home video game systems Photo camera applications

Face recognition consists of three steps [74]: detection, extraction, and recognition. Many algorithms have been written for face detection and recognition. For face detection several methods are used: knowledge-based, feature invariant, template matching, appearance based, agent based. Most effective methods are based on machine learning. Different algorithms have different precision values, that is why it is very important to choose the most appropriate for the specific use case.

In this chapter, algorithms for face detection, marks selection, recognition are reviewed. How and which information can be collected from the face is analyzed. A more detailed review of the face detection algorithm—Histogram of Oriented Gradient (HOG), Haar Feature-Based Cascade Classifier, convolutional neural networks (CNN), and face detectors comparison.

Face detection is a difficult task for computer even thought to a young child it is easy. The computer needs exact instructions [54] and restrictions and these processes require many computational resources. Images consist of pixels, which is written as binary code that's why it can not be analyzed just from discrete pixels. Face pose, illumination, expression, occlusion cause more challenges. These factors [10] have a huge impact on algorithm performance. Pose variation: the

best position is frontal but in the images can contain many faces with large pose variation. That impacts algorithm performance. Illuminations: shadows on the face, camera, and environmental conditions can decrease image quality and its fitness for recognition. Face expression: face mimics can change the distances and alignment of facial features. Occlusions: hats, glasses, beards can hide traits and complicate image pre-processing. Therefore face detection consists of three steps: pre-processing, classification, localization. First of all variability in the faces is reduced, frontal view images are used as a positive control. Standard algorithms [42] correct images lightening. After that [74] images are classified as faces and non-faces. Faces are localized in a bounding box by neural network.

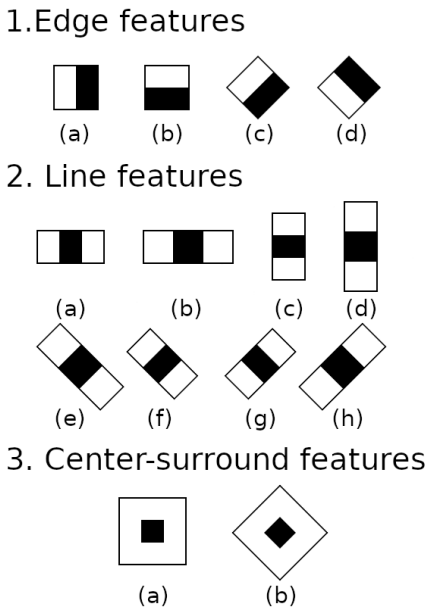


Figure 1. Haar-like features

Many machine learning algorithms are used for face detection [54]: Support Vector Machine (SVM), Hidden Markov Model (HMM), Neural Network, AdaBoost, etc. Most progressive are AdaBoost and SVM. Face detection became popular in commercial products [82] when P. Viola and M. Jones described the technique for feature selection based on AdaBoost. That algorithm does not need many computational resources and has a relatively good face detection score. The algorithm is used in OpenCV, an open-source BSD-licensed library. Haar Feature-based Cascade classifier uses different boosting techniques: Discrete Adaboost, Real Adaboost, Gentle Adaboost, and Logitboost. Cascade classifier consists of several simpler classifiers and as a result of combination is classifier with good accuracy and short computational time. To capture features from images are used rectangles in different sizes and shapes, which called Haar-like features (see Figure 1). All faces of humans have some similarities which can be associated with Haar features. The Haar features [57] can be defined as different values of pixels in individual rectangular areas. These can be any size, any position. This creates invariance from face size in images. The algorithm is based on darker and lighter zones in a human face, for example, eye sockets are darker than cheeks or nose (see Figure 2). Consequently, the face can be detected based on nose, eyes, lips size and location. Although, the feature-based system is faster [82] than the pixel-based. But feature-based also has defects such as image should be with frontal face and light in the picture has a strong impact, shadows can decrease algorithm accuracy.

One of the most popular face detection algorithms is HOG (Histogram of Oriented Gradient),

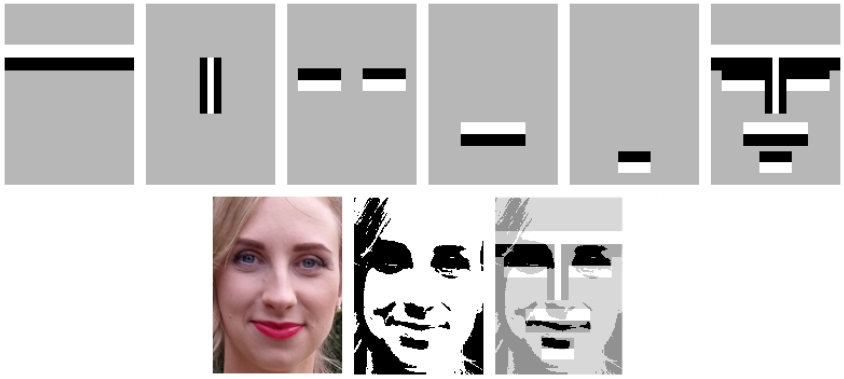


Figure 2. Adaboost learning algorithm operating principles

which uses linear SVM (Support Vector Machine). Feature descriptors' popularity increased after N. Dalal and B. Triggs presented it at the Conference on Computer Vision and Pattern Recognition [17]. The researchers first applied this algorithm to detect various positions of the human body and lighting conditions in photos. Principle of operation of HOG algorithm is based on the fact that the appearance or shape of objects can be fairly accurately 81 characterized relying on the distribution of gradient intensities or the direction of the contour lines. The algorithm is implemented by splitting the image into many small spatial regions ("cells") that are plotted on a histogram, describing the direction of gradient and intensity of the change (see Figure 3).

The direction is normalized by contrast, thus eliminating its influence. The algorithm is good at being resilient for changing environmental conditions that affect the contrast and illumination of the images. First of all, the picture is converted into gray-scale [58] for optimizing computing resources. In the next phase, the images are analyzed by pixel zones, and gradient vectors replace this zone. The vector represents the intense growth of the pixels in the picture. The gradient is a scalar vector that is oriented in the direction of increasing values. Increasing vales are lighter zones. Gradients are commonly used for edge detection in images. Further, after evaluating the directions and values of the gradient vectors, the image is split into overlapping areas. Overlap ensures consistency of values even if the outline is at the edge of the image segment. Finally, values are normalized in each block, and a histogram of block values is generated. For object/non-object classification linear Support Vector Machine is used. HOG feature vectors of the face [17] are used to detect the most similar zones of gradients. SVMs look for the best hyperplane for data separation. The purpose of SVM is to select the best categorizing vectors. SVM has some advantages [71] such as it is effective in high dimensional spaces, memory-efficient, works well when existing notable separation between classes. The disadvantages of Support Vector Machine are that they are susceptible to noise. If there are single elements which cannot be distinguished in the line, it is better to ignore such elements than to try to classify them. Another critical aspect that in reality often data is not linearly separable. Hyperplane can not be used in those situations, mathematical transformations (called kernels) correct the data before it submit for classification.

The cascade face detector based on AdaBoost and Haar-like features [36] has excellent performance in real-time. However, in real-life perfect conditions can not be assured that all images would be with frontal faces, and that is why the detector efficiency is not good enough. Also, cascade structures need high computational costs and an extra step—training stage. Even so, convolutional neural networks (CNNs) based on cascade architecture are a better fit for real-world



Figure 3. HOG scalar vectors. The gradient is a scalar vector that is oriented in the direction of increasing values. Increasing vales are lighter zones.

applications. Different face orientations, expressions, occlusions, illumination do not have such a significant impact on positive predictions. These advantages come from CNN's ability to learn new features automatically analyzed a considerable amount of training data. In the article "A Convolutional Neural Network Cascade for Face Detection" [44] the authors describe the pipeline of the detector. At the first stage with 12-net scans and 12-calibration-net reject more than 92.7 % of false predictions. Concluded image pyramid consists of a face covered 12x12 kernel detection windows. After that, resized images into 24x24 kernel and used it on 24-net. Overlapped detection windows are eliminated with non-maximum suppression (NMS) and 24-calibration-net. And at last bounding box around detected faces is a result of the procession with 48-net and 48-calibration-net. Although, detector deals with a significant visual variation, filters with a lack of the diversity of weights can limit CNN.

All of those detectors have some advantages and disadvantages. Advantages of detectors can be seen in Table 2. Haar Cascade and HOG run in real-time on CPU, but HOG does not detect small faces, and the formatted bounding box does not cover the entire face. But HOG is better for the detection of non-frontal or faces with small occlusion. CNN compared to the two other, is the best in non-frontal faces detection, and easily get hold of with occlusions. HOG does not detect looking down or up faces of people. Vikas Gupta compared concrete detectors [26]: Haar Cascade and deep neural network (DNN) from OpenCV and HOG with CNN from Dlib. Comparing the speed of face detection from 300x300 px image and not used GPU best result belong to HOG (79 FPS) and worst—CNN (4.5 FPS), but CNN outperforms others (310 FPS) if GPU is used. In another experiment with non-frontal faces best performance has neural networks; they detect not only turned face to the side as HOG but even faces, which are looking down or are oblique. It looks like a neural network is a better approach than others, but CNN fails when image resolution is low, or faces are small. Which detector should be used depends on face detection conditions, for static frontal pictures, such as in employee authentication system, Haar detector could be used. HOG could be used for live stream video processing.

Table 2. Comparison of face detectors. Symbol "+" mark that detector stands out as an advantage, "-" score that does not have that advantage, "+/-" under certain conditions has this advantage.

Advantages	Haar Cascade	HOG	CNN
Runs at real-time on CPU	+	+	-
Detects faces across various scales	+	-	-
Works for different face orientations	-	+/-	+
Robust to occlusion	-	+/-	+
All face fit to the bounding box	+	-	-

1.2 Cyberattacks which are directly related to the human being

Cyberattack by CISCO: "is a malicious and deliberate attempt by an individual or organization to breach the information system of another individual or organization. Usually, the attacker seeks some type of benefit from disrupting the victim's network" [53]. Unfortunately, the number of victims of cyberattacks and the damage to society is increasing. A number of security breaches are rising. Compared to 2013 and 2018 [45], the number of security breaches increased by 67 %, and the average cost of cybercrime increased by 72 %. The average cost of cybercrime in 2018 was US$13.0 Million for each company. During attacks, business is disrupted, valuable information and revenue are lost, equipment is damaged. Most losses are due to malware, Web-based attacks, Denial-of-service. In Table 3 describes the consequences of each cyberattack. The heatmap shows how every attack makes the most prominent loss. The most significant losses are due to the loss of information.

Table 3. Consequences of different types of the cyberattacks, values are in millions of US dollars [45].

Attack	Business disruption	Information loss	Revenue loss	Equipment damage
Malware	0.5	1.4	0.6	0.1
Web-based attacks	0.3	1.4	0.6	-
Denial of service	1.1	0.2	0.4	0.1
Phishing and social engineering	0.4	0.7	0.3	-
Malicious code	0.2	0.9	0.2	-
Ransomware	0.2	0.3	0.1	0.1
Malicious code	0.1	0.2	0.1	-

The victims of cyberattacks are not only companies but also individuals. The FBI's Internet Crime Complaints Center (IC3) as every year submit an annual report [56], in which are described that 351,937 complaints of Internet crime was registered and users had US$2.7 Billion losses in 2018 across the globe. Phishing/Vishing/Smishing/Pharming is one of the most successful attacks which affect people. According to the report, one of the most vulnerable groups is older people over 60.

1.2.1 Cyberattacks

Cyberattacks exploit human and computer vulnerabilities. This is usually a multi-step process. Lockheed Martin described seven stages of cyberattacks [33] and named it a Cyber kill chain. First of all, the adversary identifies the target, collect information, which can be a possible attack vector (Reconnaissance). The second stage is weaponization—preparation for the operation. Malicious email, USB stick left in someplace, compromised websites are several delivery choices. At a fourth stage, code is executing in the victim's system, taking advantage of known vulnerabilities (Exploitation). After exploitation, the malware installs in the asset and attacker can remotely control the victim system. And at the last stage—Action on Objectives—the adversary achieve his goal: collect necessary data, destroy the system, etc.

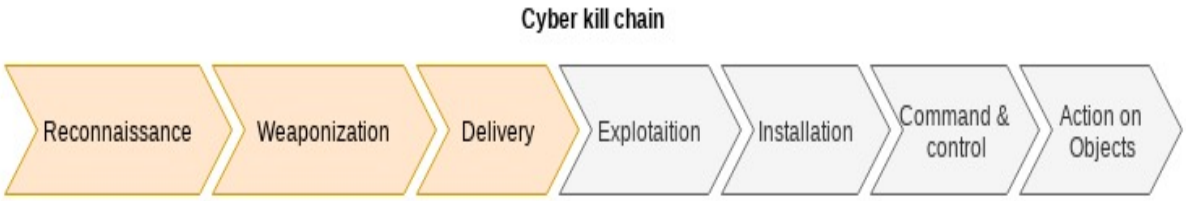


Figure 4. Cyber kill chain by [33]. Interaction with humans involves the first three phases.

The success of any cyberattack depends on the preparation stage. Interaction with humans involves the first three phases of the Lockheed Martin kill chain: reconnaissance, weaponization, delivery, see in Figure 4. This interaction includes the type of cyberattack—social engineering. "Social engineering—a means of gathering information for an attack by relying on the weaknesses of individuals" by Mark Ciampa [14]. Social engineering is a mental and emotional manipulation that leads to action or disclosure of information. During social engineering, the hacker is gaining the victim's confidence to manipulate emotions such as compassion, willingness to help, fear, greed, curiosity, anger, friendship, patriotism, vanity, altruism, belonging to a community, sense of duty and authority. It forces the victim to make impulsive decisions rather than rational ones [12]. There are a few principles behind which social engineering is successful: authority, intimidation, consensus/social proof, scarcity, urgency, familiarity/liking, trust. Authority—pretending to be a higher ranking person, or quoting a higher-ranking person. Intimidation—threatening such as if people do not do what is being said to them, they will be reported to the authorities. Social proof - people do something if someone else has done it before. If a colleague has done something, so he also must do that. Scarcity—when the lack of something creates the urge to take some action. Urgency—the rush to do something for lack of time. Liking—a person will do something if he likes another person, so subservience, finding common interest is a way to influence a person. Trust—confidence. These are the principles with which human contact is made, one of the initial steps for the cyberattack.

An attacker manipulates a person's weaknesses, thus creating conditions for accessing sensitive information, luring money, and harming a person or business. After gathering information about a person, other types of attacks is possible to achieve the goal, such as watering hole attacks, spam, hoaxes, baiting and phishing (see Figure 5). The watering hole attack targets a specific small group of people who share the same interest and visit the same website. The site is infected with malware, so computers of people in a targeted group also become infected and so the hacker can collect specific information coming from these group IP addresses. Spam—unsolicited emails. It is estimated that 9 out of 10 emails are spam. Moreover, even though only 0.001 % respond to

spam and buy a product, it still allows making a lot of money for spammers, because they can send 6 million or more emails. The problem of spam is not only that it is annoying, the person acquires unnecessary things, but also much time is spent improperly, reading and deleting emails. Estimates that in the USA, every company loses $ 874 per person due to spam reading as it reduces employee productivity. Hoax is a message that warns of the threat of a non-existent computer virus. Usually, the message is a chain email. A letter is instructing recipients to forward it to anyone they know. Baiting is an attack that is successful in 45 % of cases, according to a study conducted on the University of Illinois campus [64]. USB, CDs containing malware are left in certain visible places during the attack. They can be with some companies logos, special intriguing headers. The person is tempted to see what is inside the storage medium, and the malware gets on the victim's computer.

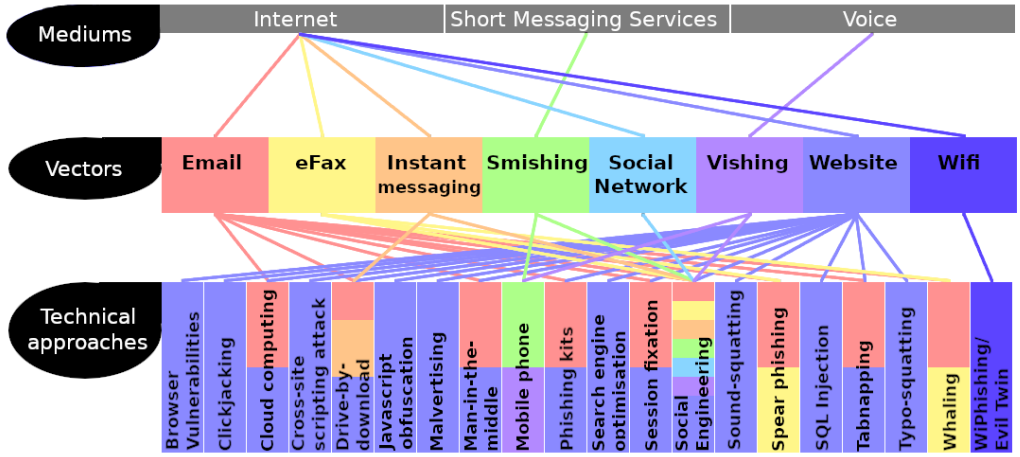


Figure 5. Phishing techniques and their mediums and vectors [12]

The most successful attack is phishing by impersonating a legal company and sending an email or redirect to the fake web page to lure private information like personal data, bank codes, social security numbers. Phishing is so successful because of its presentation to the user, on the web page, or in the emails are used the companies color schemes, logos, structure, and that is why it is difficult for people to distinguish that it is a hoax. There are several types of phishing attacks: vishing, spear-phishing, smishing, whaling, pharming. Vishing—voice phishing—phone calls used to retrieve information from people. Spear-phishing individual emails attack that is carefully designed to get a single recipient to respond. Smishing is another type of phishing attack where SMS messages are used to expose a person to specific information or to force him to click on a malicious link. Whaling—attacks on high-profile, wealthy individuals, attacks are more attentively planned, and the aim is to lure large sums of money. Pharming is a cyberattack designed to redirect site traffic to another, spoofed site. Pharming can be done by modifying the master file on the victim's computer or by exploiting vulnerabilities in the DNS server software [14].

In Lithuania, with a population of 2.7 million, in 2018, 250 thousand cases of social engineering were recorded [11]. There has been a 25 % increase in people-related cyber incidents over the year, which occurs because too little attention is paid to educating people about cybersecurity. Sig-

nificant investments are made in infrastructure protection, but into people too small. Unfortunately, in the event of a cyber incident, the software, the competencies of the employees, or a user, who lacks knowledge of how to manage cybersecurity risks, is blamed.

1.2.2 Human traits that can be used for cyberattacks

The human is the weak point in cyberattacks. 95 % of attacks are directed to personal computers [9], which are more vulnerable and not in such a good condition as companies computers, which are in professional care. People share too much personal information on internet sites, which can be used in cyberattacks. Based on statistics and results of human engineering, a hacker can make a better strategy for an invasion of the computer used by a particular person. It was determined that women and older people have worse cybersecurity habits than others. Also, is the correlation between the area of interest and vulnerability. As described in the article "Correlating human traits and cybersecurity behavior intentions" [24] humanities students create weak passwords; it depends on the interest in technologies and cybersecurity. And in certain situations, many people tend to share their passwords with their loved ones, colleagues, or friends.

Poor cyber hygiene, poor cybersecurity practices lead to potential security gaps. Therefore, potential risks related to possible traits must be assessed to determine the appropriate risk score. Cybersecurity behavior can be assessed in four categories: general demographic traits between people, individual characteristics, risk-taking preferences, decision-making style. When studying different age groups in terms of cybersecurity, it would seem that older people should be more susceptible to cyberattacks because of more poor cyber hygiene. However, in many articles [24, 50, 75, 23] indicate that younger people are more vulnerable to phishing attacks, the age group 18–25 is indicated as the worst in protecting their passwords and computer updates. The possibility of these results is that younger people spend more time in the network environment, they lack cybersecurity knowledge, and most do not have experience with cyber incidents. Such results may not fully reflect reality, as the research environment is most often a university (students and staff). Therefore, most older employees have higher education, have more knowledge, or were exposed to cyberattacks [23]. Of course, the aspect that older people are more cautious and suspicious (even 3.69 times) is a consequence of their experience; this aspect should be characteristic of the majority of the population. As a result, they are more receptive to cybersecurity training due to better impact assessment if good cybersecurity hygiene is not followed [24]. Because the research environment is a university, there is a lack of data on younger age groups than 18 years. However, the results would probably be similar to or worse for then the 18–25 age group, also due to a lack of experience and knowledge.

It has been observed that older people are more pessimistic due to their negative experiences and lower health prospects [59]. However, a study [85] done with USA and China citizens shows that cultural things also determine pessimism or optimism. The article describes that older Americans are more optimistic than younger, and older Chinese are more pessimistic. Research of privacy settings for social network usages [40] determined that older people have lower privacy settings. But they also use social networks less, reveal less information about themselves, their circle of friends is narrower, and strangers are less likely to be accepted into their circle of friends. Younger people are more likely to socialize and look for new acquaintances, so potentially young people try to protect their privacy by setting their own security settings.

Women are more vulnerable in the sense of cyberattacks [60], has lower-level computer skills, often lack knowledge of cybersecurity, and succumb to feelings in certain situations. Neuroticism

and loneliness increase the risk of cyberattacks because people are more likely to use social networks, become more addicted, for example, to the internet. Therefore, it is natural that a person who spends more time at a computer has a higher cyber risk. Due to neuroticism, it has been suggested in many articles [60, 28] that it increases the accessibility of cyberattacks. Also, scientists have unanimous opinions about the extroversion—it reduces risk. While the majority's initial hypothesis about extroversability would be that it increases risk, as such, people would seem more confident in others. However, trust in others is more of a characteristic of open people. The results obtained in this aspect of personal traits differ between different studies. Open people are not susceptible to phishing attacks, but they are more likely to disclose more information about themselves and not to secure their data on social networks.

Risk-taking is divided into five categories [24]: ethical (RTE), financial (RTF), health/safety (RTH), recreational (RTR), and social (RTS). It has been found that the more a person takes the financial risk, the stronger the passwords they use, as opposed to people who take health/safety risks. It is possible that people who take financial risks are more inclined to think about their decisions, to consider how much the consequences of a certain event will cost them, and that these people may have higher education. Decision-making style can be categorized as avoidant (DMA), dependent (DMD), intuitive (DMI), rational (DMR), and spontaneous (DMS). Multiple articles [60], [24] confirm the claim that avoidant decision-making style is negatively correlated with strong passwords creation, proactive awareness. Also, people with dependent decision-making styles have insufficient proactive awareness. People tend to make rational decisions and tend to take care of the proper security of their devices. As can be inferred, people with spontaneous decision-making styles do not have adequate safety hygiene. Based on human characteristics, a table of traits and potential risk score was compiled see Table 4.

Direct research on how ethnicity is related to cybersecurity risk assessment could not be found. Possibly not only on the question of how ethical it is to assess ethnic groups but also on whether ethnic groups can be defined in the process of acculturation, the unification of world nations. However, people with a common cultural background must have specific common characteristics; for example, the ethnic group can be more pronounced on one of the Big Five factors. Many comparative studies have been carried out with different ethnic groups living in the same country, such as Asian Americans and European Americans [19]. European Americans are more extrovert, open, conscientious, Asian Americans are more neurotic. Collectivism is inherent in Asian countries, and it is negatively correlated with extroversion and agreeableness. This cultural aspect may be one of the reasons for the lower extrovertability. However, living in a foreign community can also lead to this. It is also difficult to say that the assumptions made would be specific to the entire ethnic group of Asians; the results of people living in Asia and non-Asia can differ. Because people living in a foreign country are affected by acculturation, they take over both countries' cultural specificities; of course, this happens over several generations of people. As the article describes, a first-generation living in a country is less extroverted, more closed-minded than the next generation because it takes time to adapt to cultural change.

In the article "Group Differences in personality: meta-analyses comparing five U.S. racial groups" [21] has described not only differences between white-skin people and Asians, but also from other ethnic groups: Latino-Hispanic, Indian, Black-skin humans. White-skin people are the most stable emotionally and more extroverted, with little difference only with Latino-Hispanic. Black-skin and Latino-Hispanic are more conscientious than White-skin humans. There are almost no differences between Black-skin and Latino-Hispanic, on a contrary Black-skin and Asians are the most different groups. Black-skin people are more emotionally stable, extroverted, and consci-

entious. It is believed that there would be a more significant difference between groups if people would be analyzed from different continents, and not just within the borders of one country where nations arrived and assimilated. Of course, 40–55 % of personality traits caused by genetics last a lifetime.

The study described in "The Geographic Distribution of the Big Five personality traits" [69] shows that the assessment by ethnicity or continents sometimes is too abstract to be meaningfully correct unless people of different races living in the same country are assessed. For example, Western Europe is quite extroverted, but France, with Bangladesh, is the most introverted of the 56 evaluated nations. Of course, excluding exceptions, evaluation by region are accurate enough. The most extroverted countries are Serbia and Croatia, which is quite evident from a regional comparison, which shows that Europe's whole is quite extroverted. Compared to other countries, people living in Lithuania are more extroverted than introverted. However, according to Agreeableness, Lithuania, with Japan and Ukraine, scored the least. Most Disagreeableness people live in East Asia, as can be seen from individual country analyzes. And Agreeableness is a trait of the people of Africa, like that of the Democratic Republic of Congo citizens. Stereotypically, we would attribute a high level of Conscientiousness to East Asians. According to the study, countries such as Japan, South Korea score the least, and the Democratic Republic of Congo, Ethiopia the most. This result is possibly explained by different standards of work and interpretations of concepts. In Japan, there are high demands, and potentially people evaluate themselves very moderately in general. East Asians are also closed, as other studies have shown that people in collective countries do not show openness. And the most open people live in Chile, Belgium. While South Americans are most receptive to ideas and experiences, they are characterized by neuroticism as East Asians. The democratic republic of Congo and Slovenia has low anxiety and depression levels.

The characteristics of the Big Five traits are determined not only by the environment in which the person lives but also by genetics. It is not surprising that people from the same region where there were common ancestors had more similarities. A study of twins found that a 40–60 % variance of Big Five traits is heritable [63]. Of course, the percentage of heredity similarity is lower among relatives, but not twins, due to more significant genetic variation, but this still indicates the influence of kinship relationships on personality. The hormones whose production is driven by genes also have an impact on personality traits. Testosterone acts only on openness, but oxytocin levels do not affect only on extrovertability [61]. Thus, from the point of view of Big Five characteristics, if, for example, a person is characterized by the appearance of individual traits, it is possible to assume about the level of hormones in the body and increased or decreased hormone production leads to changes in socio-sexual orientation, disease expression. And then, based on genetics, it is possible to make assumptions about family members as well. And as mentioned earlier, family members increase the risk of cybersecurity.

1.2.3 Risk assessment

The probability that vulnerability is the risk is called the risk assessment [1], which is essential to ensure information security, as it not only identifies weaknesses and threats but also helps with preventive measures. Risk assessment consists of two parts—analysis and evaluation. First of all, is identifying the source of security risk, later evaluate different aspects, analyze, and inference. The result of the assessment is a possible impact to target, although that not guarantee full protection, but risk scenarios can minimize risk, raise cyber hygiene.

The risk describes as a set of threat scenarios, likelihood, and damage. Security risk based on

Table 4. Traits evaluation by risk scores. “+1” score if several articles have described that trait raises up risk or correlation is more significant than 0.3, “-1” score, if several articles have described that trait reduces risk or correlation, is lesser than -0.3, half a point if only in one paper wrote about that trait, which correlates with the cybersecurity aspect or correlation is low. “-” is no data or correlation. In age groups, other scores than 1 or 0.5 are used to express how the risk score increases or decreases in a specific aspect, as people get older.

Traits	Securment	Password	Awareness	Updating	Phishing	Settings	Pessimism	Addiction	Article
Woman	-	0.5	0.5	0.5	1	-	-	-	[24, 75]
18–25	-	1	0.5	0.5	1	-	-	-	[24, 50], [75, 23]
26–35	-	-	0.1	-	0.5	-0.5	-	-	[50, 23], [40]
36–45	-	-	0.2	-	-	-0.5	-	-	[50, 23], [40]
46–55	-	-	0.3	-	-	-	-	-	[50, 23], [40]
56–65	-	-	0.4	-	-	-	-	-	[50, 23], [40]
66+	-	-	0.5	-	-	0.5	-	-	[50, 23], [40]
Agreeableness	-0.5	-	-	-	-	0.5	-0.5	-	[24, 28], [77]
Neuroticism	-	-	-	-	1	0.5	-	1	[24, 28]
Openness	-	-	-	-	1	0.5	0.5	-0.5	[24, 28]
Extroversion	-0.5	-	-	-	-1	-	-	-0.5	[24, 28]
Conscientiousness	-0.5	-0.5	-	-0.5	-	-	0.5	-0.5	[24, 28], [77]
RTE	-	0.5	1	0.5	-	-	-	-	[24, 20]
RTF	-	-0.5	-	-	-	-	-	-	[24, 20]
RTH	-	0.5	0,5	1	-	-	-	-	[24, 20]
RTR	-	-	-	-	-	-	-	-	[24, 20]
RTS	-	-0.5	-	-	-	-	-	-	[24, 20]
DMA	0.5	1	1	0.5	-	-	-	-	[24, 20]
DMD	-	-	1	-	-	-	-	-	[24, 20]
DMI	-	-	-	-	-	-	-	-	[24, 20]
DMR	-0.5	-0.5	-1	-1	-	-	-	-	[24, 20]
DMS	-	-	-	-	-	-	-	-	[24, 20]

generic formula:

$$Risk = \sum_{i \in \mathcal{T}} p(i) \times g(i) \times c(i) \tag{1.1}$$

where

p—probability of threat;

g—genetically determined vulnerability;

c—consequences;

$\mathcal{T}$—the number of classified genetic scenarios.

A risk assessment which depends on human is very complex, and many additional factors can impact results. Some human behaviors actions can be predicted from his genetic-based data such as disorders, syndromes. For example, schizophrenia—mental illness appear as intermediate hallucions, delusions, correlates with Big Five personality traits [78]: neuroticism (two genetic loci) and openness (six genetic loci). Based on these traits can be predicted that a person with schizophrenia could have a higher score of social engineering attack score. But sometimes schizophrenia can be described as a paranoid personality disorder that's why human is suspicious and can be not receptive to cyberattack. A person who does not trust others has a lower level of risk [29] and will not tend to share information even with loved ones; otherwise, the risk level is rising with each member of the family. Gender, age, education, experience also have an impact on susceptibility for cyberattacks.

Although the risk assessment in the context of cybersecurity most often deals with the impact of hardware, software, environmental, the weak point is the human factor. Risk assessment consists of many components, so D. Henshel et al. [31] first distinguish between two groups: "trust"—the human factor, "confidence"—all other factors independent of the person. The article compiles a defender trust model that is appropriate for any person facing a cyberattack. At the first level, such factors as inherent and situational are distinguished, the formation of which is independent of the person. Still, the person with experience or knowledge can tilt the situation in the desired direction. Therefore, the inherent characteristic is distinguished into behavioral and knowledge/skills. Behavioral characteristics depend on the factor of human rationality. More rational people make more informed decisions, but factors such as stress affect their actions as well. Another factor is malevolence, which also depends on environmental factors, even a calm person who is provoked by bad intentions can behave maliciously. However, behavioral factors are offset by knowledge/skills that are underpinned by teaching experience. If a person has already been in such a situation, he will already know how to behave and not give in to stress. Training also has a significant impact on behavior; for example, officers are significantly more rational in stressful situations. The characteristics of knowledge/skills are further divided into expertise and attention. Competence consists of worked years, proficiency, education, training. The more experience a human has in the cyberattacks area, the more he tends not to fall into an undesirable situation or respond quickly to it. Knowledge of your weakness, in companies simulated social engineering attack training has a positive impact on the resilience of cyberattacks. Attention is divided into situational awareness and attentional capacity. Situational awareness is the ability to notice small things, changes, unusual phenomena. The earlier it is observed that the situation is suspicious, the more the further consequences can be avoided. Attentional capacity is defined as the ability to maintain attention for as long as possible. Although different levels of attention naturally characterize people, it is a trait that can be learned. All of the trust model components mentioned above are used in risk assessment that is relevant to people. The more information from social networks, photos can be gathered about a person, the better human ability to avoid cyberattacks can be assessed.

The data's quality and accuracy are crucial to drawing the right conclusions for risk assessment. False conclusions about a person can not just offend, but lead to negative consequences if the sexual orientation or disease aspect is touched upon. Statements collected from social networks, accelerometer, and gyroscope data should be obtained for a more extended period so that no conclusions can be drawn from instantaneous events. For example, if the data were taken only at a certain point in time when the person was getting an ankle and squatting, it does not mean that he is always squatting and has a problem with his feet. Or if there is a single record of self-destruction in a person's social network, whether it can not be the conclusion that human tends self-harm. Image quality and a properly trained system are essential for risk assessment using facial recognition. Perfect photos would be if there were faces in them the frontal position, colorful, all pictures will be made from the same distance from humans, a look without occlusions, shadows, or too much lightening. The distance between the person taking a photograph and the person posing would be a particularly important aspect to assess the proportions of the face, the symmetry from which one can conclude certain diseases or ethnicity. If neural networks are used for face recognition, color photographs must be used for training, with people of different ethnicities, so that the distribution of people of different nationalities is approximate.

The neutral facial expression, like standard photo resolution, improves the likelihood of face detection and other feature detection. In low-resolution photos, the face may be undetectable or detected in the wrong places. And such cases distort not only the results of the recognition system but also the accuracy of certain traits recognition. Therefore, it is essential to evaluate the impact of photo quality on face detection and trait recognition. According to ISO/IEC-19794-5 biometric data interchange formats factors: resolution, contrast, head rotation, background, illumination, eyes, glasses, mouth, head position, color, exposure, focus. Some of these factors, such as contrast, brightness, illumination, can be expressed as formulas. In the article "Design and evaluation of photometric image quality measures for effective face recognition" [8] a calculation of face image quality index (FQI) based on Gaussian was proposed in Formula 1.2.

$$FQI = \sqrt[5]{\prod_{j \in \{C,B,F,S,I\}} G(j)} \tag{1.2}$$

where

G—the Gaussian model for non-linear normalisation;

C—contrast (color intensity difference);

B—brightness (is the average of RGB to HSB converted and normalized colors);

F—focus (Manhattan distance of image gradient);

S—sharpness (degree of clarity);

I—illumination (the sum of the intensity values with weights, which comes from the figure divided into 16 parts).

However, for a full assessment of the photos' quality is a lack of face position, photo compression, and the presence of occlusion parameters. Neural network training also uses a large number of pictures and creates the need for an automated process. Various algorithms for photo quality

evaluation are being developed. It has been found that selecting only good quality photos with such algorithms improves the accuracy of face recognition by 15–18 % [41]. If all the images in the collection are essential, by classifying the pictures with such algorithms according to quality, it is possible to use slower-performing, but better, face recognition algorithms when dealing with lower quality images. For example, Haar Cascade detector is good in detection face in various scales, CNN is better choice for occlusions and face in different pose detection.

According to the 2018 NIST report [25], since 2013, the algorithm for face recognition accuracy has improved by about 20 times. The failure rate in 2013 was about 4 %, and in 2018 about 0.2 % when studying a 1.6 M photo. The reduction in the error rate is due to a change in the old algorithm to CNN, which well recognizes faces that are not frontal pose. Neural networks such as ResNet, Inception, Special transformers have emerged. However, if twins are not even monozygotic, but the same gender, it remains a problem for face recognition algorithms, most twins are identified as the same person. Although the number of twins born is not large, the twin birth rate is about 30 twins per 1,000 births [49], but the birth rate of twins is 76 % higher than in 1980. The percentage of cases where twins are born should continue to increase, as this is related to childbirth in the late age of women and the increase in artificial insemination. Identification of twins can be improved with the recognition of separate traits, such as scars, moles, face proportions, but such identification requires good quality photographs. Therefore, the risk assessment should also take into account the error factor of the features identified by the machine training algorithms. Of course, if it is possible to compare automatically determined traits with genetic testing results, this factor no longer makes sense because genetic testing readings are invariably reliable.

2 Model

The number of cyberattacks is growing and, while companies are trying to protect themselves, there are loopholes in systems and new ways to access humans or systems. All cyberattacks are characterized by a sequence of actions defined by the term cyber kill chain. Knowing the stage of an attack can help the defender identify which stage is the current attack, calculate the probabilities of how much time is left before the end of the attack, or how long one step can last. It is observed that the attacks are repeated and repeated to achieve the desired goal [32]. A cycle is formed, during which certain stages can be skipped, interrupted, or terminated, see Figure 6. As a great deal of knowledge has been gathered about cyberattacks, defense tools and methods are being improved, defenders notice a computer attack in time and often stop it. However, it is difficult to control when it comes to social engineering when human vulnerability is exploited. For example, an interrupted attack can be resumed by collecting information about the person responsible for protecting the system or accessing it. Human vulnerabilities can be exploited at any time during an attack. Man is associated with the delivery stage and other stages where a cyberattack goes unnoticed in time.

In reconnaissance, information about the company, system, and people are collected using various techniques: port-scanning, dumpster-diving, shoulder surfing, etc. Then the attack tools, methods are collected: Trojan horses, viruses, worms, etc. The person can also be necessary for the delivery phase. Poor cyber hygiene can leave gaps in computer equipment, out-of-date antivirus software, or a person can be a vector that will open a malicious email letter, page, install a program on a computer, open a found CD or USB key. Suppose the attack is interrupted in this phase, back to social engineering, when a new victim is possibly selected [76]. In that case, information about the person, schedule, motivation, and role in the company is collected. Applied technical methods like phishing, water holing, pharming; social: pretend to be a high-ranking person, quid pro quo, intimidation; physical: dumpster diving, listening, shoulder surfing. And using the accumulated knowledge of how to distract a person, encourage him to make a certain decision, the attack is continued. Such an insert of social engineering can actually be carried out after each unsuccessful stage. And so it shows how much influence a person has in cyberattacks.

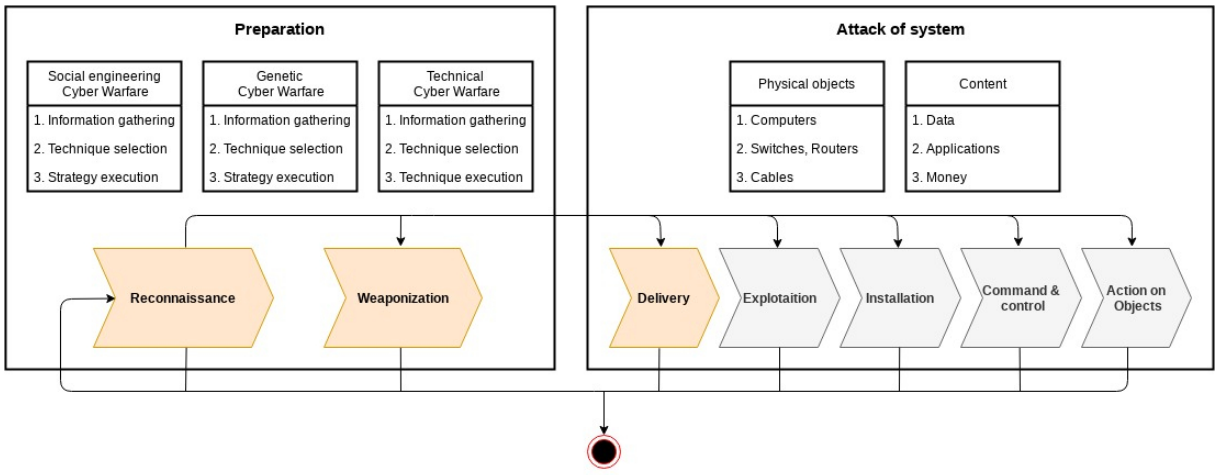


Figure 6. Cyclic social engineering kill chain

2.1 Ontology-based model

Semantic network $\mathcal{S}$ is a global network of anthologies $\mathcal{O}$.

Ontology is a collection of triplets t. All triplets t are in the set $\mathcal{T}$. Triplets can also have [62]: URI/IRI links—**U**, literals in text or numbers—**L**, or empty blank nodes—**B**.

Definition 1. Triplet Let us have $\mathcal{T}$—a set of triplets, where each triplet is in depends at least one ontology $\mathcal{T} \in \mathcal{O}$.

$$t ::= \{(s,p,o) | s \in \{\mathbf{U} \cup \mathbf{B}\} \wedge p \in \{\mathbf{U}\} \wedge o \in \{\mathbf{U} \cup \mathbf{B} \cup \mathbf{L}\}\}$$

Here s—subject, p—predicate, o—object.

In general, Triplet constructs a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a node type mapping function $\gamma : \{s \cup o\} \rightarrow \mathcal{V}$ and edge type mapping function $\theta : p \rightarrow \mathcal{E}$. Here $\mathcal{V}$ and $\mathcal{E}$ denotes the vertices and edge sets respectively and $|\mathcal{V}| > 0$ and $\mathcal{E} > 0$. Such a graph we name ontology schema.

Predicates p are the literals of `isA`, `has`, `describe`, `implements`, `depend`, `trigger`, `partOf`, `influence`, `threatProbability`, `risk`, and `property`. The Table 5 contains examples of predicates, the left side shows the predicates that can be used in the semantic network, and the right side shows an example of the use of a predicate to form a triplet.

Table 5. Examples of predicates usage relating to path encoded in ontology

Predicate	Usages example
`isA`	Head —is a→ Body parts
`has`	Head —has→ Face
`describe`	Circadian phenotype —describe→ Lifestyle
`implements`	Age —implements→ Algo/CNN/1
`depend`	Eye color —depend→ Genetic
`trigger`	Enviroment —trigger→ Psichological condition
`partOf`	Eyes —part of→ Face
`influence`	Age —influence→ Human
`threatProbability`	Openness —threatProbability→ 1.5
`risk`	Human —risk→ 4.5
`property`	Hair —property→ Curly

Definition 2. Schema definition *graph* Let us assume $\mathcal{G}$ is from triplets $t \in \mathcal{T}$. Graphs has vertices

s and o and vertices p, therefore Resource defined graph can be defined as:

$$\begin{aligned}\mathcal{G} ::= &\forall t_i(s_i, p_i, o_i), t_j(s_j, p_j, o_j) \exists t_k(s_k, p_k, o_k) | t_i \neq t_j \wedge t_k \neq t_i \wedge t_k \neq t_j, \\ &t_k(s_k, p_k, o_k) = ((o_i, *, s_j) \vee (o_j, *, s_i) \vee (s_j, *, o_i) \vee (s_i, *, o_j)) \wedge \\ &t_i, t_j, t_k \in \mathcal{T}.\end{aligned}$$

Definition 3. A property path $\mathcal{P}$ is a path that is defined on the graph schema $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and is denoted as a $v_1 \xrightarrow{p_1} v_2 \xrightarrow{p_2} \ldots \xrightarrow{p_{k-1}} v_k$.

An example of the property path is any variant of (node, property, node) variations connected in a sequential path.

Definition 4. Likelihood of a threat l is the function $\mathcal{V} \to \mathbb{R}$ producing value if the existence of a node v_i or the existence of a path $\mathcal{P}$ defined from node v_i to v_j over p_{ij} as `threatProbability` is true. The function is defined as:

$$l(v_i) ::= v_j \iff \exists v_i, v_j : (v_i, \texttt{threatProbability}, v_j) \in \mathcal{P}.$$

To simplify notation $l(v_i)$ in short is denoted as l_i. In some cases, only the existence of traits is already a threat in itself. In other cases, external factors are required to trigger the threat, which may be described in the path.

Definition 5. Impact of a vulnerability is the function $\mathcal{V} \to \mathbb{R}$ producing value if the existence of a node v_i or the existence of a path $\mathcal{P}$ defined from node v_i to v_j over p_{ij} as `impact`.

$$g_i(v_i) ::= v_j \iff \exists v_i, v_j : (v_i, \texttt{impact}, v_j) \in \mathcal{P}.$$

To simplify notation $g(v_i)$ in short is denoted as g_i.

Definition 6. Risk $\mathcal{R}_\mathcal{P}$ **on a Property Path** is a function that given a property path $\mathcal{P}$, $l \in \mathbb{R}$—likelihood of threat, $g \in \mathbb{R}$—property-bounded vulnerability impact, $c \in \mathbb{R}$—consequences, $\mathcal{P}$—the number of property paths. Then, the risk $\mathcal{R}_\mathcal{P}$ is

$$\mathcal{R}_\mathcal{P} ::= \prod_{v_i \in \mathcal{P}} l_i g_i c_i. \tag{2.1}$$

c_i is a consequence multiplier for a probable impact if the vulnerability is found. Consequence multiplier acts as importance factor to highlight property path.

The risk function is only preliminary for model computation and will be adjusted later.

2.2 Meta-schema of traits network

Modern society provides a wealth of information about itself through public networks: photos, speeches. Devices such as mobile phones, smartwatches tracking the whereabouts of the man, his physical condition. At first glance, pictures can tell race, age, gender. Still, more things can be predicted, such as if a person belongs to the east Asian ethnic group, he can not drink too much because this ethnic group has an inheritance trait of aldehyde dehydrogenase deficiency. With the help of machine learning, it is possible to predict not only such obvious things but also certain diseases or even sexual orientation. Furthermore, the analysis of comments on social networks

with machine learning can predict whether a person has an eating disorder, depression, or post-traumatic stress disorder (see Figure 39).

Anthropometric data such as gender, age, body complexity, ethnicity, facial features are quite well studied, and their recognition with machine learning is used as a commercialized product. The company "Visage Technologies" is offering the product for gender recognition, which can be used for purposely targeted advertising. Personalized shopping for best-fitting clothing is a real possibility. For online shopping applications such as "Sizerme" can be used, which take photos with multiple poses of the person in real-time. A deep learning algorithm calculates body measurements of the person from photos. It is also expected that 3D body scanners will be available in clothing stores in the near future, which will also help to find the right size of clothes. Convolutional neural networks (CNN) are frequently used to identify gender and age in photos [27, 67, 43]. However, gender is also recognized by voice [3] or gait [35]. Accelerometer and gyroscope sensors of a smartphone collect signals of person gait, a modified 1D HOG descriptor is used for data processing. Age and gender recognition is essential for digital forensics [4]. Digital data can be evidence of crime and cybercrime. Facial recognition with computer vision is critical in some cases when is needing fast analysis and decisions, such as child abuse investigations. Like human fingerprints, human face marks (moles, scars freckles, birthmarks) are individual [2] and even allow the separation of monozygotic twins. Facial marks, gender, age enable recognition of an individual, face features such as skin color, eye, nose, mouth shape, and hair distinguished from the overall complexion reflect the characteristics of a particular population. In the article "Automatic classification of human facial features based on their appearance" [22] represent facial taxonomy based on traits of races. After face features extraction, they are categorized with eigenfaces and clustered with the K-Means algorithm. Individual facial features such as hair geometry, the color of hair, lip, eye, eyebrow, skin tone [7] are also studied. Eyes analyzed with more segmentation [18]: iris, pupil, sclera segmentation, eye color. Eyes have information about individual traits: health, age, but also can describe features specific to some part of the population such as possibilities, for example if people have light eyes and skin, they have the higher risk of skin cancer. Hair shape, color, if left unchanged, could also be one of the ethnic division features. For features extractions and identifications machine learning techniques are used: one of the most popular HOG with SVM, Haar to detect and localize face, for landmark features detection libraries such as dlib. Machine learning techniques can describe external anthropometric human features; however, they can not determine from photos of innate traits, which can only be identified in genetics laboratories. Plastic surgery can rejuvenate, change gender, contact lenses change eye colors, even makeup and hairstyle can raise confusion for gender recognition. Although voice timbre, gait, thoughts enunciation are more difficult to change, they requires constant self-monitoring until it becomes a habit.

Algorithms detect some traits better than humans, such as sexual orientation from the face [83]. Accuracy for this feature by deep neural networks (DNN) is 91 % for men and 83 % for women recognition when the accuracy of guessing people is between 55–65 %. The research confirmed previous studies that there is a positive correlation between gay male face with feminity and gay female face masculinity. That is based on the difference of androgens levels during a woman's pregnancy period. Androgens are responsible for gender face traits expressions, sexual behavior. That is one of the impacting factors for sexual orientation and why it is recognizable from the face. Although tolerance in society is increasing; however, gays have a higher propensity for depression, substance abuse. This part of the population is more vulnerable, especially when national authorities are interested in citizens' lifestyle, there are countries, such as Saudi Arabia, where sentences of death are executed for same-gender sexual orientation. Therefore, it is vital that inappropriate

information does not fall into the wrong hands, as well as ensuring cybersecurity for such people.

From social media, speeches with machine learning algorithms can be recognized a person's psychological condition such as depression, PTSD, eating disorders, stress, or propensity to commit suicide. Almost 23 % of the world's myriads are due to mental health or substance use [51]. However, the psychological condition is not so easy to recognize; creating a system suitable for a large group of people is difficult simply because of language differences. Also, social networks use new terms, misspelled words, extra characters. Although it may be argued that the human psychological state is an individual matter, certain disorders occur in specific segments of the population. For example, the United States faces the problem of armed attacks at school, colleges [68], mass shootings in public areas. As a result, the part of this country's population is exposed to stress, depression. Lithuania is one of the East European countries, which has the same historical experiences as their neighbors, now the economic situation is similar between the countries of the region, however many years rate of suicides in this country was much higher than in the others.

Human physical activity is a characteristic that depends on many factors, such as body composition, age, gender, physical and psychological state, social environment. Smartphones, smartwatches have gyroscopes and accelerometers, which analyze not only the frequency of physical activity but also its nature—walking, climbing. Also, hand position dependent on small motor movements such as typing, handwriting, eating different meals, dribbling, brushing teeth [84], or smoking [15] are analyzed. Knowing the frequency of human activity and what it is, it is possible to observe habits, to guess the specific of work, to predict when and what a person will do. If the data obtained can be used to identify gait that is of a certain gender, then it should be possible to determine the state of health, age. Therefore, it is necessary to analyze the connections; for example, if a person has started smoking more frequently at some point, the human may be under stress at that moment. As people become concerned about their low levels of physical activity, which increases the likelihood of heart disease, obesity, the widespread use apps which counting steps and has led to a trend to wear smartwatches. Due to the extensive use of devices, a large amount of data is collected, analyzed with mostly DNN or CNN. Human activity is not genetically determined, but defined by a combination of many traits.

This section discusses many human characteristics that can be automatically recognized. Traits semantic network is a graph that illustrates the relationship between parts of the human body and properties, see Figure 7. The semantic network describes elementary things from the point of view, but a classification of properties is formed. The structure shows that a person has body parts; one part is the head. The head is characterized not only by external features such as eyes color, shape, but also by the human psychological state. The psychological condition also affects the person's physical activity; physical activity is related to physical performance such as strength, flexibility, speed. Thus, a schematic network is a basis on which algorithms should be selected to identify certain human characteristics. If there is a human face in the photo, it is unnecessary to identify features related to other parts of the body. The semantic network defines that case; it shows that algorithms, which identifying marks, skin color, the proportions of facial, should be used. With the emergence of new traits that can be identified by machine learning, neural networks, the semantic network can be further expanded by simply adding new traits to the existing structure in a concrete place.

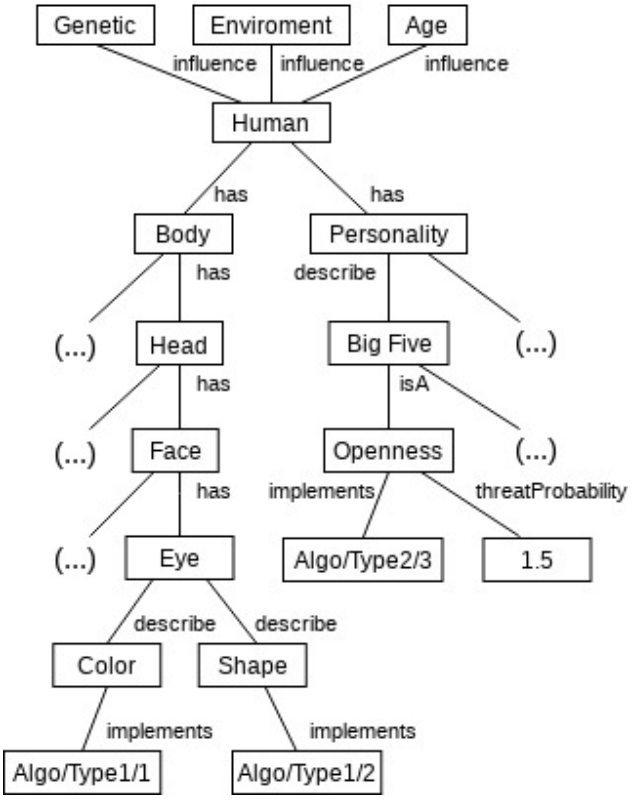


Figure 7. Illustration of part of semantic network

3 Implementation

3.1 Architecture

Django—Python web framework was used to develop the web application. Solutions implemented in this framework: site maps, admin module, security against such security vulnerabilities as SQL injections, clickjacking. Also, the Django framework makes it easy to implement a visual part that receives information from the user, using a form. Forms are used in HTML templates, and Django handles several aspects: prepare data for rendering, create HTML forms, and process data received from the client. Thus, the project was implemented using Python, HTML, CSS, and JavaScript. With Python implemented Backend, and with other languages—visual part. WebcamJS library written in JavaScript was used to take a photo using a browser. This library makes it possible to take a picture with the user's camera and sends it to the URI specified in the form. Therefore, the user does not need to have a photograph or his computer—to perform a risk assessment based on identifying the person's traits from the image. In Figure 8 shows the part of the questionnaire where also is the photo taken with a user computer camera.

The objects of prototyping system evaluation are two: photos and questionnaire questions (architecture of systems see Figure 9). The photo's purpose is to determine as much information as possible about the person without direct contact with them. When several projects are connected, the prototype system determines age, gender, ethnicity, and what parts of the body and face are seen in the photo, eye color, and whether the people are family members. The questionnaire aims to gather more specific information about a person which may increase cybersecurity risk. Questionnaire questions determine a person's Big Five traits, decision-making style, risk-taking preferences.

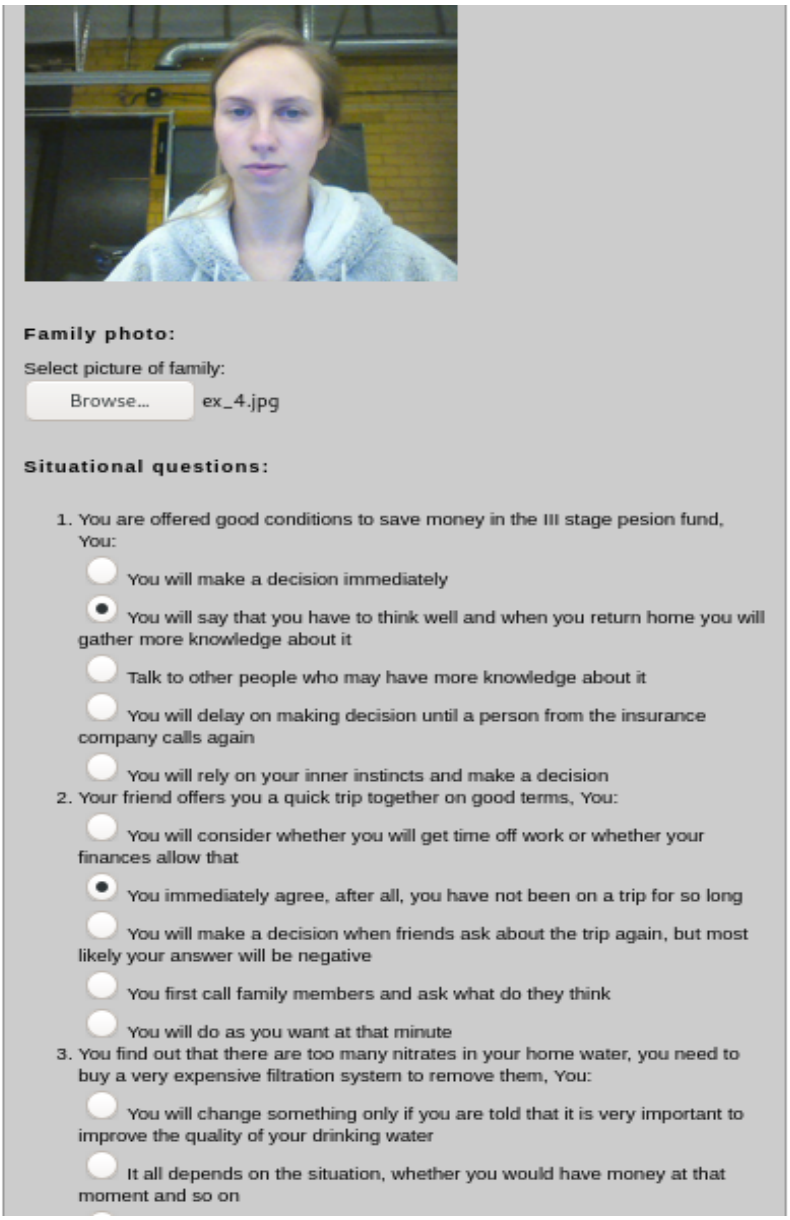


Figure 8. Example of a web application questionnaire

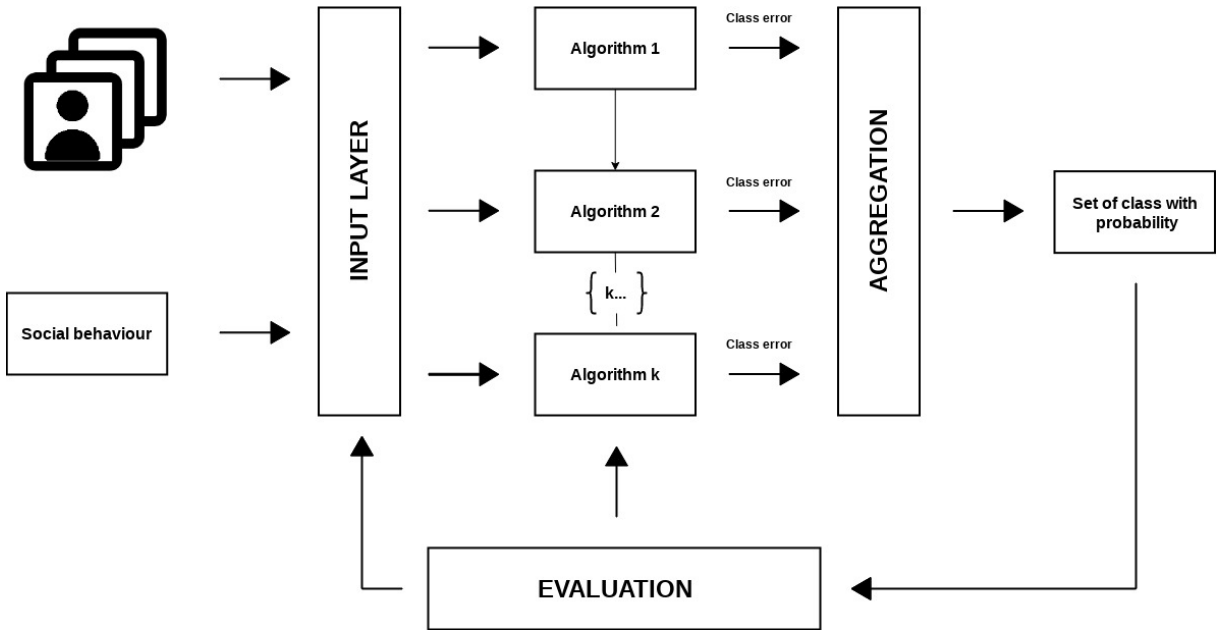


Figure 9. Architecture

The web application and MySQL database (details about the database in Section 3.3) are stored in Docker containers that contain isolated environments to run a particular application, database, or other services. Containers communicate with each other through the Bridge network when the containers have their internal IP addresses. Docker uses OS-level virtualization, which requires fewer resources than using virtual machines. The application of resources should be optimized in the future, especially to recognize more human features. The more the projects are combined, the longer the processing time of the human data. Functions should be optimized so that similar processes are run only once.

3.2 Algorithms

In the Scientific Research Project, several projects' effectiveness in identifying human traits from photographs was investigated. Each of the three features: gender, age, and ethnicity, was identified in three projects. Excluding one project with lower than 50 % ethnic recognition results, worse than results of guessing, there were four projects left: "Ethnicity classifier" [5], "AgeGender" [48], "Deepface" [73], "Age and Gender Estimation" (AGE) [80]. The projects are used to model a prototype system that will phenotypically identify cybersecurity risks.

The surveyed people who submitted a photo of themselves or their family had their traits identified from the image using previously tested algorithms in the research work. The studied algorithms did not show the accuracy expected, so in this work, not one algorithm is used, but several to identify traits as accurately as possible. Initially, it was intended to abandon the age and gender algorithm from the "AgeGender" project. Still, it was observed that in photos that fail to detect face with a HOG detector, but the face is detected with a cascade face detector based on Haar-like features, the algorithm well enough identifies human age and gender. The HOG did not detect the face, where it is in partial shadow, or the photos are quite pale.

Correct face detection has a significant impact on trait recognition. The section discusses generally the most commonly used and also used in this prototypical system facial detectors. Replacing photos uploading with a snapshot by a computer camera helps solving the photo quality problem.

Because most traits are recognizable using multiple neural networks, considerable computational resources are required. The prototype could be improved by optimizing some processes, such as face detection would be performed only once with one selected detector. The project uses similar neural networks, but with different training, for similar tasks, such as DNN, for recognizing age, gender, and body parts; in “Deepface” and “Ethnicity classifier” projects—neural network based on VGG. Thus, based on a semantic network, an algorithm is reached by walking through the Property path $\mathcal{P}$ through the photo’s traits. If a head is found in the picture and face is detected, then an attempt can be made to recognize the eyes’ location and color using an algorithm. The MTCNN is used for eye color recognition in the prototypical system. The Figure 7 showing part of the semantic network eye recognition algorithm in a prototype URI could be replaced by a URI leading to the MTCNN.

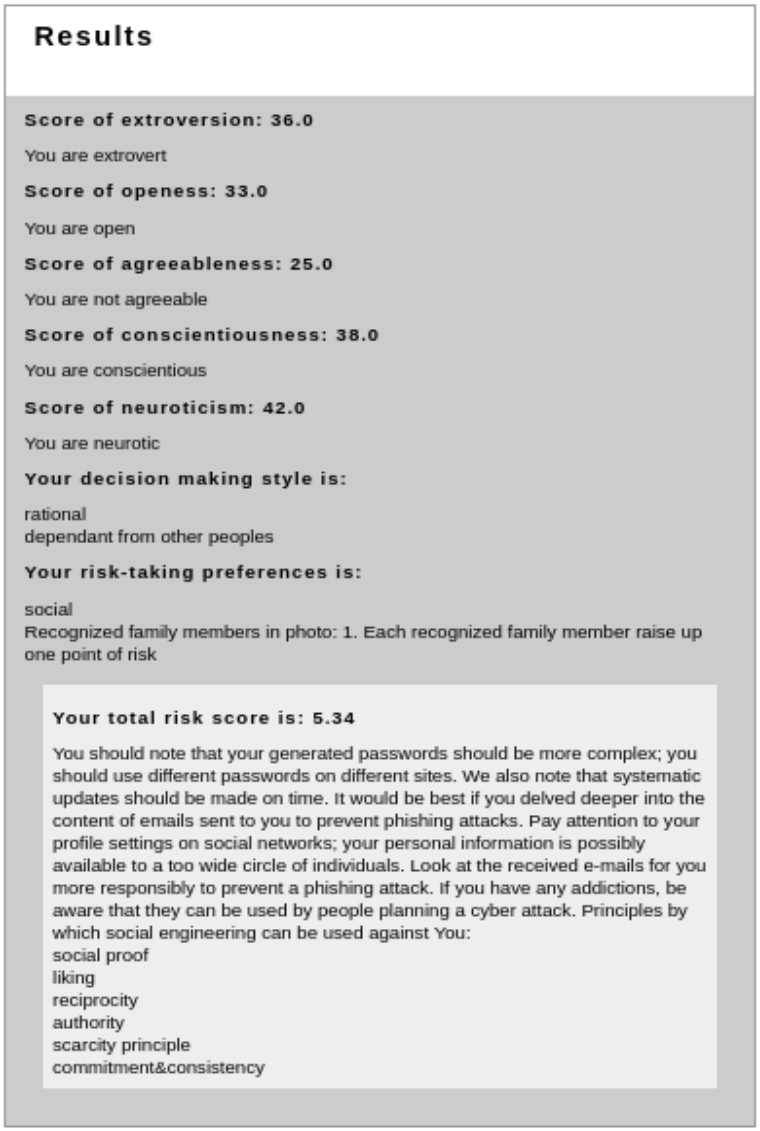


Figure 10. Example of questionnaire results

A person who would like to assess their cybersecurity risk visits the page to take a photo using their computer camera, answers the questions to identify Big Five characteristics, decision-making style, and risk-taking preferences. In Figure 10 results were obtained evaluating answers to ques-

tions and identifying features from the photos. For the risk score from the answers to the questions, the calculation is used an algorithm based on each answer's points and the risk score of each characteristic. If the family picture is uploaded to evaluate how much the risk increases if kinship is identified, the risk value increases accordingly for each recognizable family member. A retrained FaceNet from the "Deepface" project is used to identify family members. A retrained neural network architecture is a Siamese neural network where two input vectors are used with the same weights. The Figure 11 is a Siamese neural network architecture that summarizes architecture, which fits the four neural networks, attempted to retrain. The results of the trained networks and why FaceNet from "Deepface" was selected for the task are described in the Section 4.1.

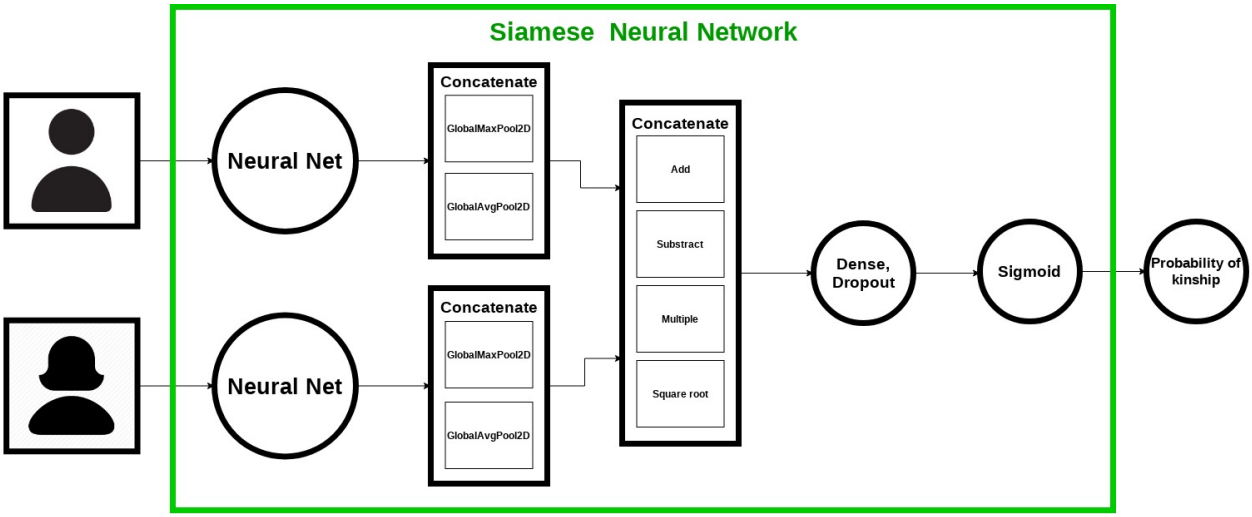


Figure 11. Architecture of Siamese Neural Network for kinship recognition

3.3 Database

The project has an opportunity to assess not only age, gender but also eye color, ethnicity, what parts of the face or body are visible in the photo. Due to some lack of resources, such as the data, ethnicity's risks are not assessed in this work. Works such as linking the proportions of the human face or other parameters to genetics are left for the future. Young age and the female gender are known as an increased risk factors from the literature reviewed. The data is stored without leaving a connection with a real person in a MySQL database about a person who will assess his risk. In Figure 12 shows a schematic of the stored data models. Not all stored fields are represented in the user model, as the stored data may change in the future. Photos of people are not saved; only the obtained results are held, such as age, gender, how much is recognized in the image of family members, and it is also planned to save the parameters of face proportions later. By not storing photos, identity data provides a secure way to store data that can be used for further research. The database also stores scores from evaluated questions, which identify human character traits according to the Big Five, decision-making styles, risk-taking preferences.

The future aim would be to tell all the information about a person only from photographs, such a system needs more research, which can be used to make assumptions about not only the physical traits of a person but also the character of a person. In the article "The Genetic Links Between the Big Five Personality Traits and General Interest Domains" [38] it was found that women are more neurotic and agreeable compared to men, and they also have a higher interest in Domestic-Manual, Social-Educational and Hedonistic-Enterprising. And men are more technical, Logic-Minded,

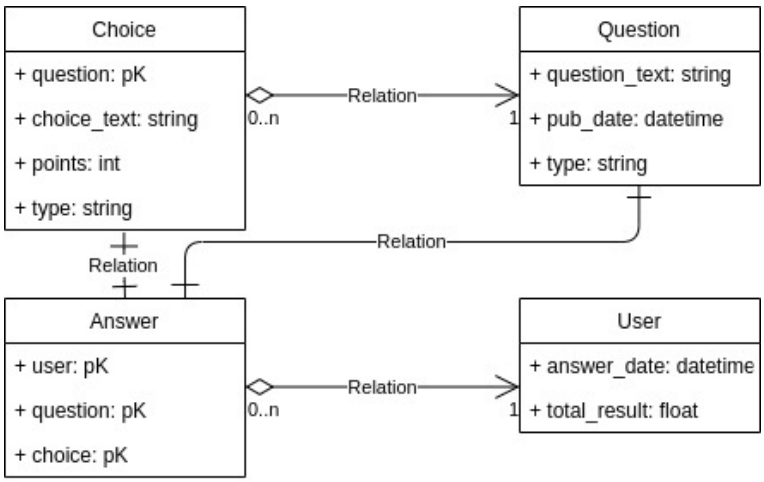


Figure 12. Models of web application

and more interested in research and science. Young people also have more Technical-Logical thinking. They are also more hedonistic, more open, and more neurotic than older people. And the older ones are more agreeable, conscientious; they have a domestic, Cultural-Intellectual, Social-Educational interest. Based on the associations of such external traits with Big Five traits, one could make a direct guess only from the photos of the risk dependent on these character traits. Thus, having both photographs and personality traits identified from the survey questions, it is possible to conclude whether the pictures would suffice for a generalized identification of at least the Big Five characteristics.

4 Experiments

4.1 Kinship identification

The task of identifying the blood ties associated with the members of the family is relevant in various fields: an investigation of crimes, search of gone missing people, identification on social networks, analysis of historical and genealogical data. From any other area of multimedia: audio recordings, gait data, texts on social networks, kinship cannot be identified, only from the photos that best show the physiological traits that reveal the common genetics. First of all, automated kinship recognition requires data for training. The data set, which consists of about 1000 families with several photos by the same family member, was created by Northeastern SMILE Lab [66]. There are people of different ethnicities in the images, but Caucasian predominate about 64 %, Latino, Asian, African, about 10 %. As this work assesses the risk of people living in Lithuania, all of whom belong to the Caucasian ethnic group, such an inter-ethnic relationship should certainly not be a problem. Images size 224x224 px. The research team, which compiled a collection of photos, also tried several different CNN training options, one of which was the best SphereFace. The accuracy of this CNN is 69.18 $\pm$ 3.68 %. ResNet was better in recognizing the connection between the grandparent and the child, with accuracy of about 65 %. As might be expected, the lowest accuracy of kinship connection recognition is between grandparents and children. The researchers managed to achieve an accuracy of $\sim$70 % between parents and children, and even slightly higher between brothers and sisters. The best identifiable kinship is between sisters, about 77 %. Appearance differences due to age, such as wrinkles, baldness, etc., have a major impact on

kinship recognition. The closer the age of comparison is, the simpler the recognition task, which becomes similar to identifying a person based on the smallest distance difference. However, if we try to align the peoples' ages in the photographs artificially, we can distort the true traits. That's why the task of kinship identification is much more complex than person identification.

Kinship recognition is essential for a person's cybersecurity risk assessment. As mentioned earlier, each family member increases each other members of their family's cyber risk score. Most attempts to recognize kinship are made based on a transfer learning methodology using VGG-Face. In this work, the "Deepface" FaceNet model was retrained to take advantage of existing projects. The "Deepface" project has six facial recognition models, with FaceNet described as one of the best. In 2015, Google proposed a facial recognition model [70] based on a deep neural network architecture that excels at the high-quality mapping between the photo and Euclidean space. The output of the model is the Euclidean distance, which directly reflects the face similarity. If the distance is small, then the photos are the same or a very similar person, if large—a different person. This model is also particular because it uses the triplet loss function, which requires three pictures: the main is called anchor, positive—which contains a person similar to the one in the main image, and the negative one—an unlike person. The function aims to make the Euclidean distance of similar people smaller than that of different people. One of the positive sides is that good efficiency is achieved using just 128 bytes per face in a photo.

In this work four models were trained, using the FIW dataset. There are 470 families in the training dataset, of which 10 % are allocated for the validation process. Equal amounts of kinship and non-kinship relations are used for training; photos are preprocessed before using in models. To train models a computer with AMD Ryzen 7 2700 Eight-Core Processor 3.20 GHz and 32.0 GB installed memory (RAM) was used. Two of them were based on FaceNet from the "Deepface" project. The first was trained by the transfer learning method; the second was retrained. The other two models based on Exception were retrained; they differ in the optimizer used—Adam was used in one case, RMSprop in another.

Training CNNs from scratch requires a tremendous amount of data. Therefore, in the case of a limited amount of data, transfer learning is applied when a trained network's features are taken over, and the system is trained to solve new similar problems. FaceNet from "Deepface" was trained to identify the similarity. It is a close task to recognizing kinship. It just needs more parameters for family members recognition; it is not enough to determine just similarity distance. Facenet from the "Deepface" project was also retrained in this way to recognize kinship. All FaceNet layers to the fourth from the end were freezed, and the last four (Classification, Dense) set as trainable. Feature extraction layers were added to perform fine-tuning. Transfer learning requires less training data and less parameters that should be updated therefore it required less computing resources and less time. Using transfer learning FaceNet model accuracy did not achieve the expected results; from the first training epoch to the last, accuracy increased by about 16 % (from $\sim$50 % to $\sim$66 %), see Figure 13, validation results were slightly worse, accuracy finally reached nearly 62 %. Overfitting was not observed because training and validation accuracy are quite similar. Training loss value decreased from $\sim$82 % to $\sim$60 %. However, the validation loss has hardly changed. The validation loss data curve is quite jagged and uneven. Fluctuation is not observed in training loss because, in each epoch, it is usually calculated based on the entire training set, but the validation loss is usually calculated in one mini-batch of the validation set in each epoch, so it has more oscillations. Therefore, to clarify the average data result, the principle of a sliding window is applied when the data is smoothing. However, in this case, even after receiving a smoothed validation loss curve, a fluctuation was still observed, which may indicate that there was too little data at validation to

summarize the performance of the model. For training and validation, same number of kinship as non-kinship relations were provided. Photos were used from the same "Recognizing kinship in the Wild" dataset, initially divided into training and validation. Images were of different quality and various facial positions.

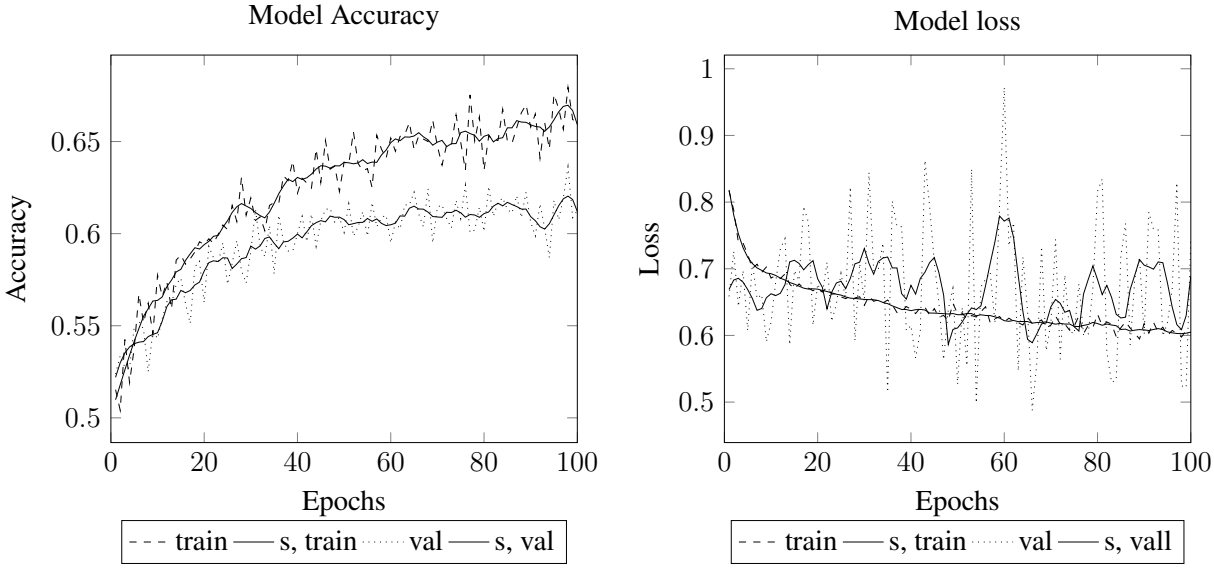


Figure 13. Metrics of transfer learned FaceNet. s—smoothed.

In this case the transfer-learning methodology has not proved successful, or more various modifications were needed, as the overall accuracy achieved is only about 63 %, see the Table 6. The model recognized the kinship of blood in different age groups with relatively similar accuracy; no uniqueness was observed that the relationship between people of similar age was better recognized. Model recognition is marked by a high recall score; high recall is when kinship is detected, when it exists. But this situation occurs because the retrained FaceNet model assumes that the relationship between humans is kinship. There is a large number of False Positive, and thus, a type I error occurs.

The transfer learning methodology was not as effective as expected, therefore in this work, FaceNet from the "Deepface" project were retrained to identified kinship. A Siamese neural network is being developed (see Figure 11). The first two layers are input layers for images, the size is 160x160 px. Therefore, photos from the kinship recognition image database "Families in the wild" (FIW) [55] before usage are processed: reduced, normalized. FaceNet layers to the third from the end are set as trainable. For usage in GlobalMaxPool2D layer, photos are reshaped. In the concatenate stage, differences in images between people are revealed by the transformation of features. Transformation of features is used based on other developers experiences, articles [55, 46]. Kinship recognition solves the binary problem if kinship is 1, if not—0. Therefore, the sigmoid function is used in the Dense output layer to obtain the result. An equal number of kinship and non-kinship connections are used to train the neural network.

The model's training accuracy increases from ∼51 % to ∼85 % and the validation from 52 % and accuracy in 61 epochs is already a well-established parameter and even somewhat seems to start decreasing, see Figure 14. From the loss plot, it can also be seen that the validation loss in the 61 epoch is one of the lowest values, and later as if the values begin to grow. Apparently, from the graph, after about 61 epoch, the training process can be completed because the model is no longer learning. The training loss line continues to shrink, with a possible overfitting effect.

The validation loss line is also jagged as in the transfer learned FaceNet model described above, possibly for the same reasons.

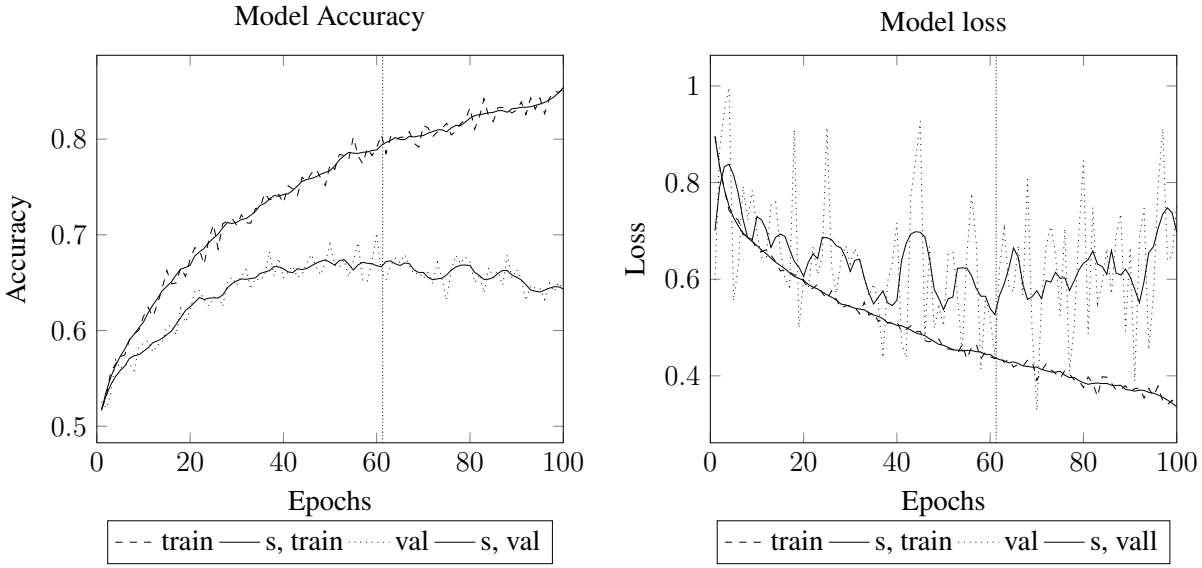


Figure 14. Metrics of retrained FaceNet. The vertical densely dotted line indicates a place where the teaching process could have been stopped. s—smoothed.

Model accuracy determined by testing with photos from FIW reaches almost 76 %. The Table 6 also shows metrics for each different kinship type. The relation between the grandparents and the grandchildren is the worst recognized in the examined literature, but FaceNet, trained in this work, recognizes these relationships no worse than the others. Even the best accuracy was determined between grandfather and granddaughter. In 74 % of cases where there is a suspicion that there is a kinship, this is a correct guess. In about 80 % of cases where a kinship had to be detected, it was detected. In most cases, inferior results were obtained when recognition was performed between members of different gender.

In the literature reviewed, retrained Facenet, VGGNet with ResNet50 are used in most cases for kinship recognition. Based on a comparison of Keras available models [39], the Xception model has better accuracy compared to VGGNets or ResNets. Xception is an improved variant of the Inception model that differs in the order of operations and in the absence of non-linearity after the first stage [13]. The model is used with 299x299 px photos; the size of FIW photos is 224x224 px, so the parameter include_top set as False is included when training the model. The model is similarly retrained as Facenet, using first time Adam optimizer and another RMSprop, which gave better results in training the neural network for facial expression recognition [47].

If we look only at the training data, everything is well modeled, the accuracy increases from $\sim$52 % to $\sim$84 %, and the loss decreases from $\sim$72 % to $\sim$35 %, see Figure 15.

However, observing the validation data, we see that the accuracy slightly decreases, and the loss results increase quite rapidly. If overfitting is observed, the accuracy may stop growing or even decrease, which is observed during this model validation. Such an increase in loss with decreasing accuracy is possible due to the cross-entropy loss classification used. When wrong guesses are punished more than the right ones are rewarded. Therefore, even if many kinships are correctly identified, one faulty connection has more weight to increase the loss. Consequently, it is possible to conclude that the model is overfitting, not suitable loss function for the model, or ideal for this classification task.

Table 6. Metrics (%) of different families members kinship recognition. Relations: brother-brother (BB), sister-sister (SS), siblings (SIBS), father-daughter (FD), father-son (FS), mother-daughter (MD), mother-son (MS), grandfather-granddaughter (GFGD), grandfather-grandson (GFGS), grandmother-granddaughter (GMGD), grandmother-grandson (GMGS). Pink indicates the best accuracy results, blue—precision, green—recall, yellow—F1 score results.

Metric	BB	SS	SIBS	FD	FS	MD	MS	GFGD	GFGS	GMGD	GMGS	General
Transfer learning—FaceNet												
Accuracy	62	61	60	62	65	63	61	63	65	66	59	62.5
Precision	58	57	57	58	60	58	58	59	60	60	55	58.2
Recall	85	83	82	86	91	87	82	89	92	97	93	87.9
F1	69	68	67	69	72	70	68	71	73	74	69	70.0
FaceNet												
Accuracy	74	79	73	76	77	77	76	80	73	76	71	75.6
Precision	74	78	72	74	75	74	75	77	71	74	68	73.8
Recall	73	81	74	80	82	83	78	85	78	81	79	79.5
F1	74	80	73	77	78	78	77	81	74	77	73	76.5
Xception+Adam												
Accuracy	75	76	74	67	68	69	68	63	70	63	62	68.6
Precision	81	79	78	74	76	75	76	70	84	69	67	75.4
Recall	67	71	65	51	53	58	53	47	50	46	48	55.4
F1	73	75	71	60	62	65	62	56	63	55	56	63.5
Xception+RMSprop												
Accuracy	77	76	73	66	69	69	66	68	65	66	63	69.5
Precision	81	81	78	75	76	77	74	78	74	75	69	76.2
Recall	69	67	64	49	55	56	48	52	46	47	45	54.4
F1	75	73	70	59	64	64	59	62	57	58	55	63.3

The situation of a model trained using Adam optimizer is similar to that of an RMSprop optimizer. Training accuracy increases from 50 % to 84 %, and loss decreases from 74 % to 35 %, see Figure 16. Validation accuracy initially increases from 54 % and reaches a maximum of 61 % at 14 epochs, and later the results begin to decrease. The loss initially diminishes and is stable until the 19 epoch and then begins to grow. Therefore, an overfitting effect may also be observed.

Xception with RMSprop gave better accuracy (69.5 %) and precision (76.2 %) results. However, Xception with Adam better detected kinship when a kinship had to be detected, true positives. Both trained models have the most difficulty in identifying the relationship between grandparents and grandchildren, and the best results are with individuals, which are of a similar age: brothers and sisters. FaceNet has much better recognition of kinship among people of different ages, both parents and grandparents. Both Xception models feature better precision results, which show that kinship is identified when it is. Also, these two models better recognized brother-brother relation than FaceNet. Retrained model Facenet is used in kinship recognition for risk assessment due to the better result shown. The trained model results in this work are better than those described by the authors of the photo dataset FIW. They are also significantly better than the human results obtained when distinguishing whether it is a kinship. Only 57.5 % of cases people are right [55]; they are especially faced with the problem of recognizing kinship among people who are of a different ethnicity than themselves. In this respect, trained neural networks do not face such a situation, and that is superior.

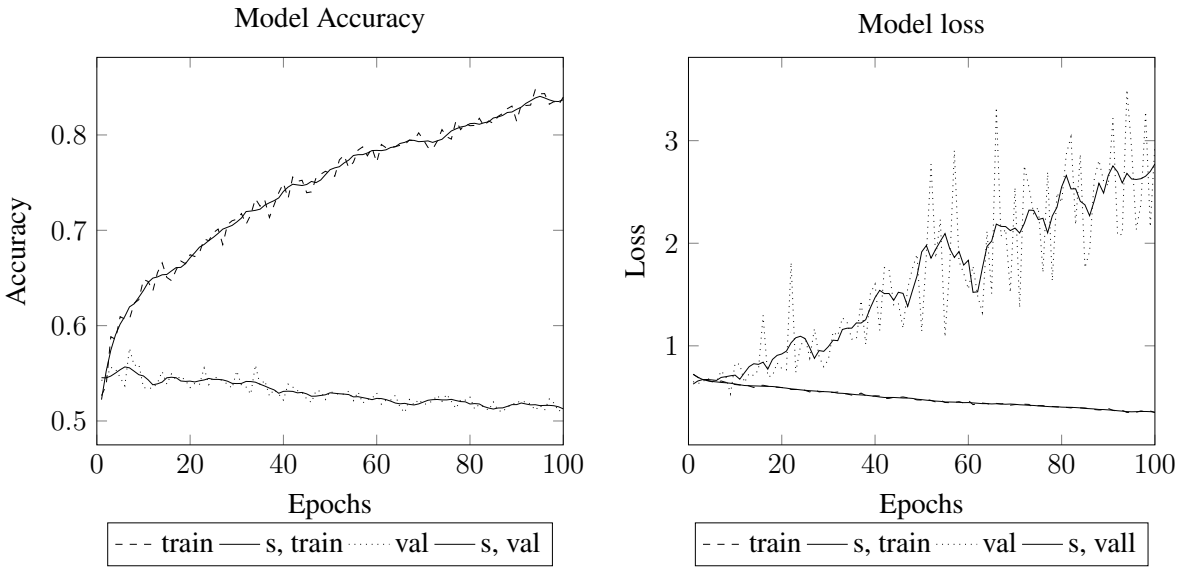


Figure 15. Metrics of retrained Exception with optimizer RMSprop. s—smoothed.

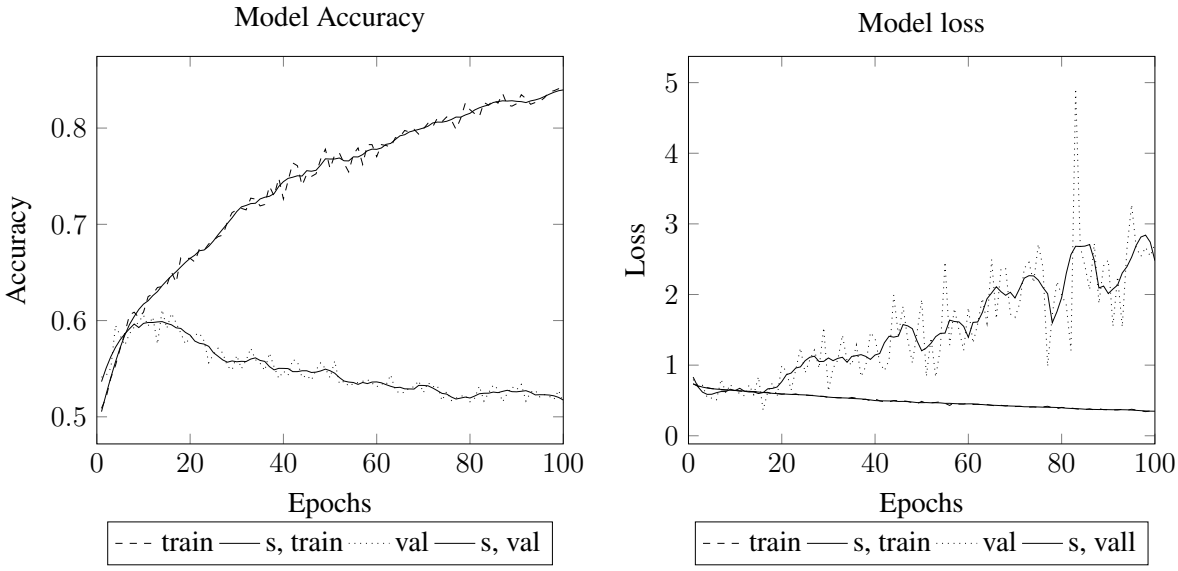


Figure 16. Metrics of retrained Exception with optimizer Adam. s—smoothed.

4.2 For risk assessment extracted traits from collected data

In this work, the people's cybersecurity risk is assessed based on the submitted photos, which determine the age, gender, ethnicity, and whether the people in the picture are in kinship. And based on the answered questionnaire questions, which determine Big Five characteristics, decision-making style, risk-taking preferences. From the analyzed literature, women and youth belonging to the 18–25 age group are considered to be at higher risk than men and other age groups. Although no direct study has been found on how ethnicity is associated with cybersecurity risks, it is possible to make assumptions based on articles analyzing Big Five's traits dependence on race. Asian is most vulnerable because they are more neurotic, introverted, not so conscientious. And all of these aspects increase the risk. White people are the least susceptible because they are more emotionally stable, more extroverted than other ethnicities. However, in this research work, only Lithuanian residents will be studied; unfortunately, most of the country's population is White ethnicities; such assumptions about ethnicities cannot be confirmed. According to a study that analyzed 56 countries' people according to the Big Five [69], Lithuanians are neurotic, not agreeable, not conscientious, but quite extroverted and open. These traits show that Lithuanians belong to a high cyber risk group based on the Big Five characteristics. Not conscientious, disagreeable people are characterized by poor cyber hygiene, and neuroticism and openness are positively correlated with the success of phishing attacks and shoddy storage of personal data on social networks.

The objects of most research is the university community: students and staff. Information technology students are often involved in cybersecurity risk assessments, so many people are engaged in research who know about cyberattacks and how to avoid them. Thus, the innovativeness of this work is the mere fact that the subjects are family or at least several members of a family. Research on family members is more complicated because it requires several members' willingness to participate in the study. Therefore, it is more informative in the genetic sense of heredity, allowing specific causes of traits to identify. Although Kinship identification from photos has been tried for several years, there are still difficulties identifying it. Because blood kinship identification relies on distance between vectors, as does face recognition, there is a problem in determining whether people are kinship when they are from different families but very similar as Will Ferrell and Chad Smith. The task's complexity is also because, due to age differences, children and grandparents, although they have common traits, are very different. In this work, there are no such high demands on recognizing family members. It is a minimal possibility that there will be a very similar person in a human environment and not a relative. As well as analyzing the human, the attacker would most often make assumptions about people in the same photos as the victim. Therefore, this work is limited to kinship determination without identifying human roles in the family.

Parents should take care of safe cyberspace, but in these times, when children use technology from an early age, the opposite is true when children teach their parents [16]. Children are one of the leading brokers of information technology innovation, and they have a particularly strong impact on older people over the age of 55. Therefore, although parents are more cautious because they have life experiences, they lose their role in information technology. As a result, there is a potential for lack of security in cyberspace and the transfer of less secure knowledge of technology. The cybersecurity risk associated with loved ones increases with each member of the family as they could expose more information about the victim of a cyberattack. Maybe, the person does not share information about himself, but other family members upload photos from vacations and favorite restaurants. The article "Open source intelligence for risk assessment" [29] is also suggested that marital status should be evaluated because it can create the possibility to blackmail the

person. For example, if a person cheats on a spouse. So any close person can make a situation for the attacker to access the person, even an uploaded photo of a pet with his name can reveal, for example, the victim's password. Women, older people, people with lower socioeconomic status, with lower levels of education have less knowledge of cybersecurity and technology, so they are more vulnerable and pose a higher cyber risk to themselves and their loved ones.

4.2.1 Social behaviour

The questionnaire was based on trait identification standards. The Big Five traits are identified using a 50-item International Pool of Personality Item Pool [34], which was compiled in collaboration with researchers seeking to have advanced personality measures. For each of the five traits, agreeableness, conscientious, extroversion, neuroticism, and openness, ten questions are used, which are negative and positive wordings to ensure the defendant's answer is warranted. Five decision-making style is: avoidant, dependent, intuitive, rational, spontaneous. The General Decision-Making Style Questionnaire (GDMS) [72] is used to determine the decision-making style. But only part of the question is used in this work questionnaire because they are all very similar. Instead of unused questions, five situations are formulated with five behavioral variants that reflect each human decision-making style. There are also five distinguished risk-taking preferences: ethical, financial, health/safety, recreational, social. The risk-taking tendency is assessed using a Domain-Specific Risk-Taking (DOSPERT) Scale consisting of 30 hearings [6]. Participants were also asked questions to determine cyber hygiene according to the Security Behavior Intentions Scale (SeBIS) [20]. These questions presented to identify if studied traits by residents of another country similarly correlated with the cyber hygiene habits defined in the Table 4.

The study involved 23 families consisting of 2 to 9 people and two individuals, altogether 87 people counting the questionnaires answered. Some families are made up of just husbands and wives, but they are not counted as kinship. More women (63.6 %) than men (36.4 %) participated in the study. Age groups from 18 years are studied; most individuals (43.2 %) aged 25–34, least aged 35–44 (4.5 %), and over 65 (5.7 %). Only 9 of the respondents are currently students, but most have completed a bachelor's or master's degree. Research people have education in different fields: social sciences, engineering, life sciences, and seven students or already graduating in information technology. However, computer literacy was marked as advanced or experienced by as many as 74.7 % of people. Of course, this self-assessment depends on the people themselves, so a critical assessment is possible. And the answers to the question of how much time you spend at the computer were distributed almost evenly after about 20 %, only 5–6 hours were marked by 12.6 % of people. Thus, the general data show that the study population is quite diverse. As the subjects are Lithuanian residents, as expected, ethnic diversity did not occur. But most of the research subjects do not belong to the university community and are family members, which makes this research unique.

The survey involved 87 people, which is the initial set needed to prove traits and risk model concepts. It cannot be said that this set describes the whole population in terms of some characteristics because more open, extroverted, agreeable people usually agree to participate in voluntary research. Of the people in the study, 64 people were identified as prone to extrovertability, i.e., scored more than 60 % of points. In almost all age categories, more than half of the people in that category rated themselves as sociable, outgoing, energetic, see in Figure 17. Although, as you can see from the Table 4, extroverts are not susceptible to phishing attacks, they are good at device security practices, they are not addictive, but they have quite a lot of information about themselves

and people around them on social networks. And this gives the attacker a lot of data for creating various attack scenarios. Extroverts often start doing things without thinking about everything, so the attacker creates situations where sudden decisions need to be made and can get the desired result. Only six people in the study did not describe themselves as open. It has been noticed that open people are particularly susceptible to phishing attacks. Not only do they reveal a lot of information on social networks, but they also often have low-security settings that allow others to access personal info. Knowing about the risks posed by traits and their compounds, it is possible to strategically train people to protect their personal information and be less trusting.

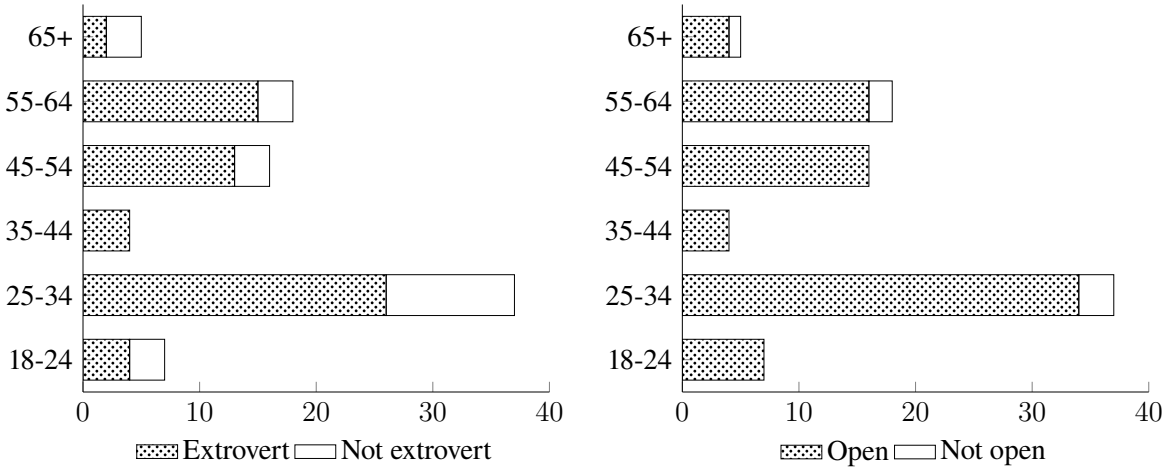


Figure 17. Extroversion and openness correlation with age.

Agreeable people have a strong willingness to help others, pleasant, they believe in other good intentions. 69 people identified as agreeable, four people belonging to the 35–44 age group also identified themselves as conscious, see Figure 18. Similar situations exist in other age groups, so some traits are often related to each other or lead to another. Agreeable people often trust others too much, thinking that the intentions of the others are good. Therefore, there are possible cyber risk scenarios where an attacker can use social engineering techniques such as pretending as high-rank persons or IT department employees to access data from agreeable people. Conscientious people are hard-working, with a clear goal, an orderly, accurate person. They have good cyber hygiene, timely updates of hardware, use passwords, or shut down computers without using them. In some respects, these people are at the lowest risk than people with other Big Five features. However, even such people have a weak point; possibly, a person uses a system that stores all their passwords, so an attacker entering that system would make it easy to access all other important information.

Of the Big Five traits, the most vulnerable people belong to the neurotic group. They are more likely to be exposed to phishing attacks because, for them, it is harder to detect where a lie is. They also do not use the right privacy settings on social networks, although they often spend a lot of time on social networks because they feel lonely. Less than half of the study participants identified themselves as neurotic, with no one in the 35–44 age group, see in Figure 19. The study involved 55 women and 32 men. 43 % of women in this work described themselves as neurotic and 34 % of men. The percentage of men was more than women in only one category—men are more agreeable; as many as 81 % of men described themselves as such. Due to the current trend: women have less competence in information technologies, generate weak passwords, do not perform timely updates of computer programs, have poor cyber hygiene, are more gullible, and are more likely to be involved in phishing attacks than men. Thus, some traits like conscientiousness may reduce a

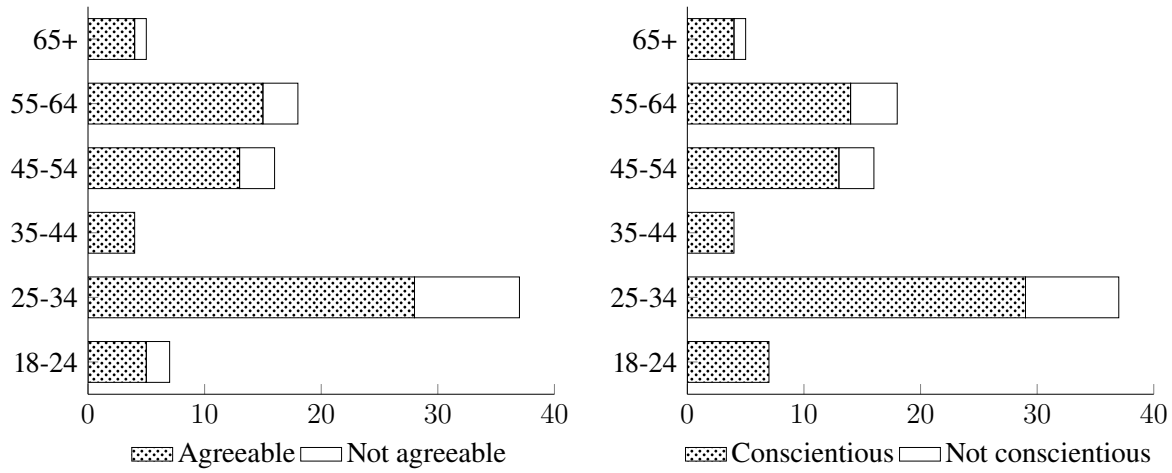
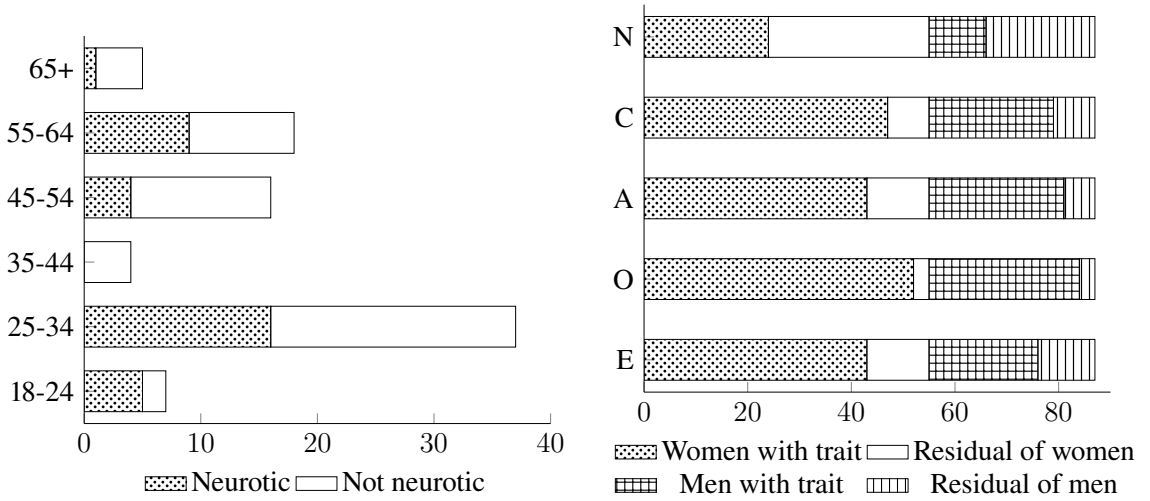


Figure 18. Agreeable and conscientiousness correlation with age.

woman's risk of a cyberattack. Still, if a woman is neurotic, she may be an easily accessible target to the attacker, so proper corporate attention for people with certain traits is incredibly essential.



(a) Neuroticism correlation with age.

(b) Big Five traits correlation with gender. N—neurotic, C—conscientious, E—extrovert, O—open, A—agreeable

Figure 19. Big Five traits correlation with gender and age.

The study of risk-taking preferences showed that Lithuanians are not inclined to take risks, especially on financial or ethical issues. Those most involved in the research are not afraid to take risks in recreational and social areas: to enjoy active, dangerous leisure activities, and are not afraid to speak up on aspects that run counter to their interests, but this is more noticeable in younger generations. Older people are reluctant to take risks on these issues because of the visibly historical legacy; the life experience impacts more aged people. Of the 11 people, six who tend to take recreational-risk belong to the 25–34 age group (see Figure 20), which already usually has the money they could spend on such entertainment. People who enjoy extreme entertainment should know that this is a theme that a person planning a cyberattack would try to start a conversation with them.

There are only two people in the study who have ethical and financial risk-taking preferences, see Figure 21. Lithuanians are quite conservative on economic issues, afraid to take risks, and there

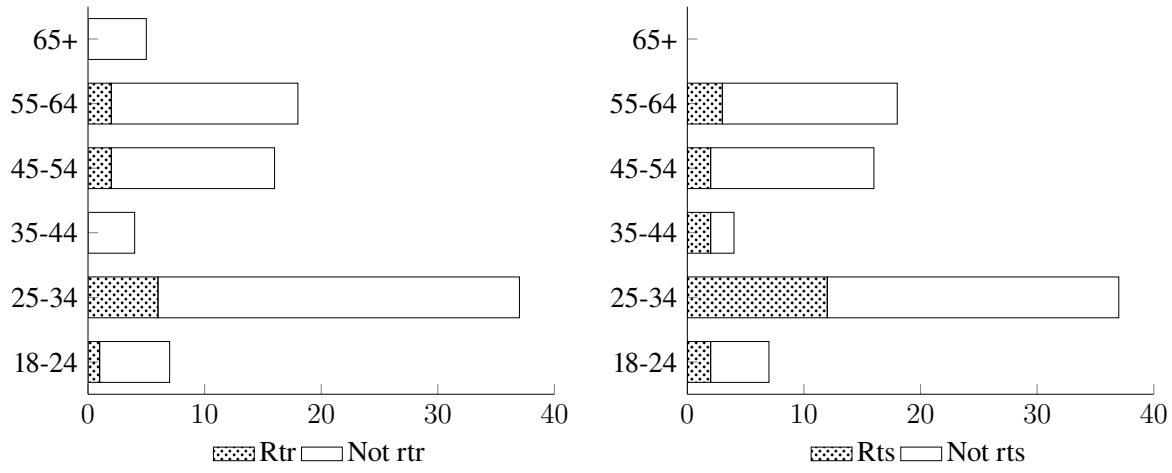


Figure 20. Risk-taking preferences: recreational and social correlation with age.

is no gambling culture in this society like in America. And of ethical issues, people usually have certain attitudes that they think they will follow, but they do not know if they will act the way they think in a real situation. People with ethical risk-taking preferences in previous studies have found that they have bad cyber hygiene habits: they do not update the programs in time, they generate weak passwords. They also often reveal others' secrets, so they pose a risk to their companies', family members' cybersecurity because they can disclose too much information.

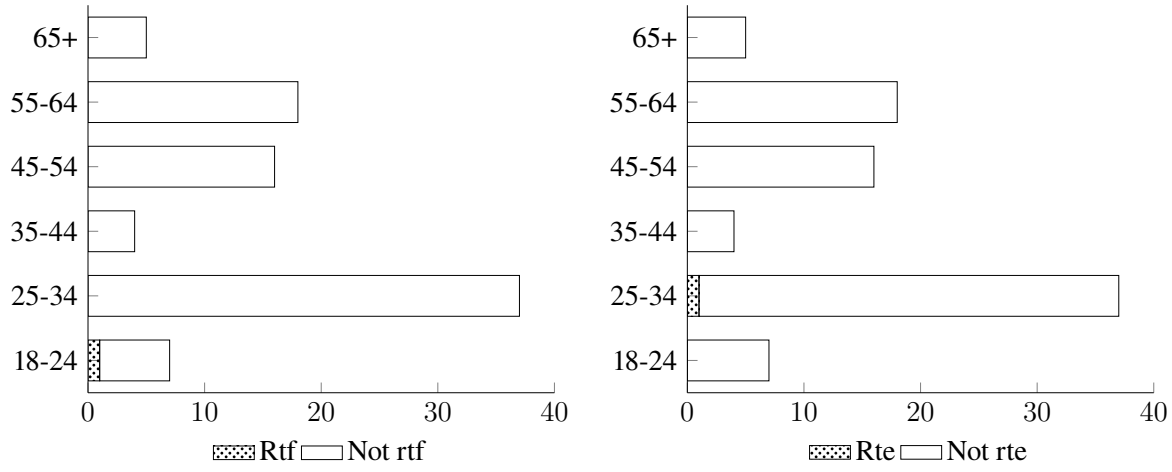
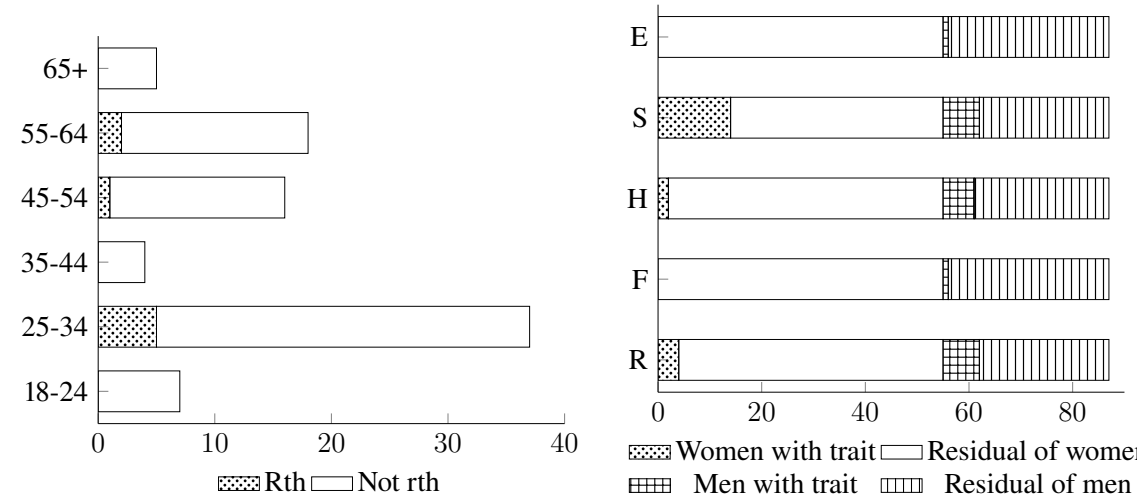


Figure 21. Risk-taking preferences: financial, ethical correlation with age.

Only eight people revealed that they have health risk-taking preferences, see Figure 22. These days, people are greatly influenced by promoting a healthy lifestyle, creating the right image in society, for example, the importance of driving with seat belts. Therefore, although people with health risk-taking preferences have a higher risk of cybersecurity due to their poor cyber hygiene habits, habits can be changed if people are adequately trained. The survey data show that men are taking more risk; women are more at risk only in the social field than men; they are not afraid to express their opinions. The older people are, the more cautious they are, the less risky. People in the 65+ age group were not inclined to take risks in any of the categories. Older people are not gullible, so they are quite difficult to access through social engineering. Their weakness often can be a lack of knowledge in information technology that can lead to security vulnerabilities in computer equipment.



(a) Health/safety risk-taking preferences correlation with age.

(b) Risk-taking preferences correlation with gender. R—recreational, F—financial, H—health/safety, S—social, E—ethical.

Figure 22. Risk-taking preferences correlation with age and gender.

Many people surveyed do not have a concrete decision-making style because it usually depends on the situation. One man has avoidant and seven people—intuitive decision-making style from the interviewees, see Figure 23. Women most often rely on intuition when making a decision. Decisions made by people who are dependent on impulse can be unpredictable as people who make decisions spontaneously. Previous research has not shown that any of these two decision-making styles are associated with risk factors, which is unfortunately not the case with people belonging to a group of avoidant decision-making style. People with an avoidant decision-making style are often not characterized by good cyber hygiene. Such people delay their decisions in time and may therefore be exposed to a cyberattack because they will not take timely action.

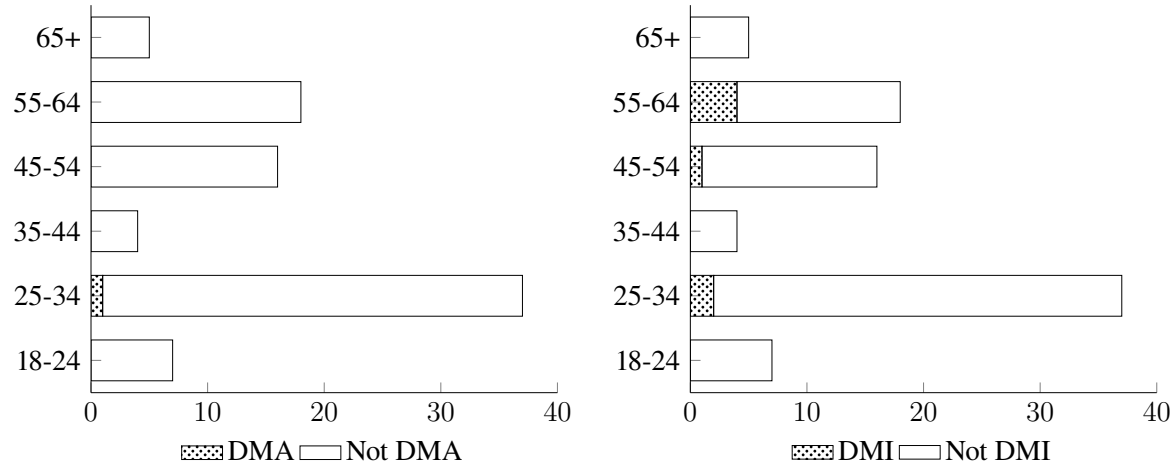


Figure 23. Decision making style: avoidant and intuitive correlation with age.

More young people (18–24, 25–34 years old) have a dependent decision-making style, see Figure 24. That situation occurs because they possibly live with their parents or are financially reliant on them. Therefore, by establishing such a type of decision-making, it is possible to make assumptions about a person's family situation, his/her relations with relatives. And although this type of decision-making does not show inferior cyber hygiene based on previous research, however,

the individual can be reached by an attacker when communicating with a person's loved ones.

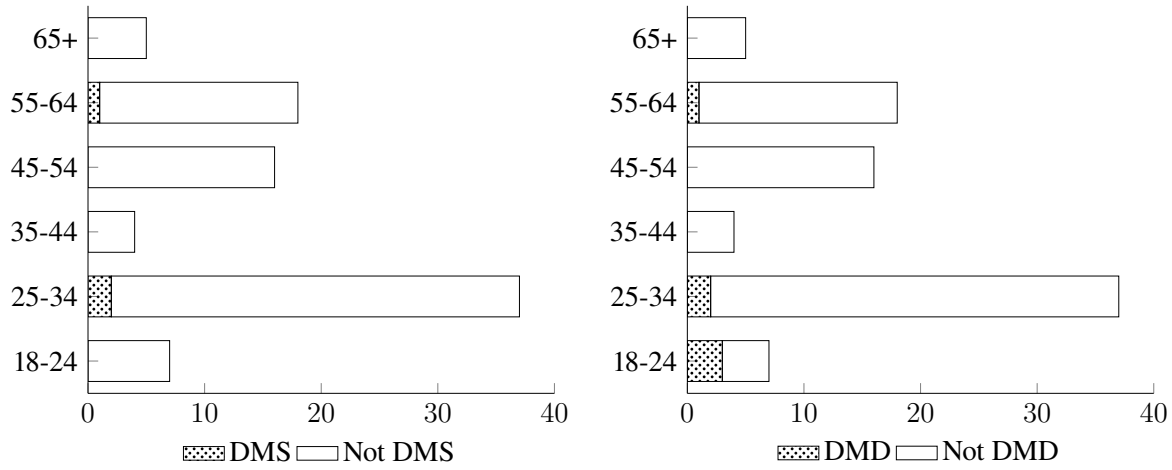


Figure 24. Decision making style: spontaneous and dependant correlation with age.

In this work, most people have a rational decision-making style—about a third of the people who took part in the study. The percentage of men with such a trait is higher ($\sim$53 % men) than women, see Figure 25. Because of their propensity to act prudently, they often have more of the necessary knowledge. In this group, people make timely updates, generate strong passwords, protect devices, and take other security measures. So this is one of the traits that reduce the risk of cyberattacks.

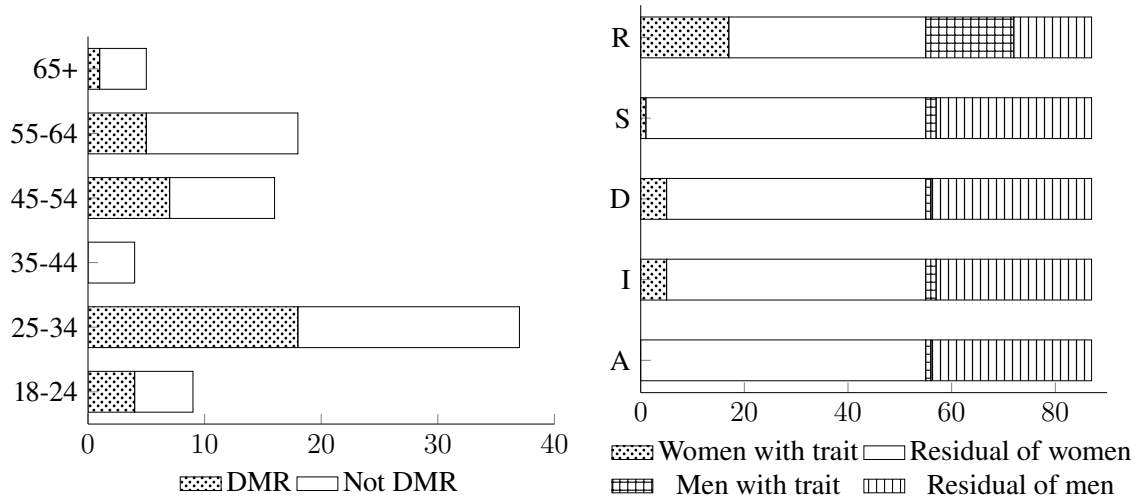


Figure 25. Rational decision making style correlation with age in the left plot and all preferences correlation with gender in right. A—avoidant, I—intuitive, D—dependant, S—spontaneous, R—rational.

In this work cyber risk score based on human traits is automatically calculated. And to assess at least partial risk validity, cybersecurity questions were asked to participants under SeBIS, as well as 14 questions to find out if a person had already been caught in a cyber or social engineering attack. For example, would a person reveal private data to a person who introduced himself as a high-ranking person, or a person tends to write the passwords on a paper. Periodic updates are executed promptly by 23 people (see Figure 26), most of them younger, more aware of the importance of updates, and better fluent in English. Half of all participants have security awareness. Also, most of them are younger people.

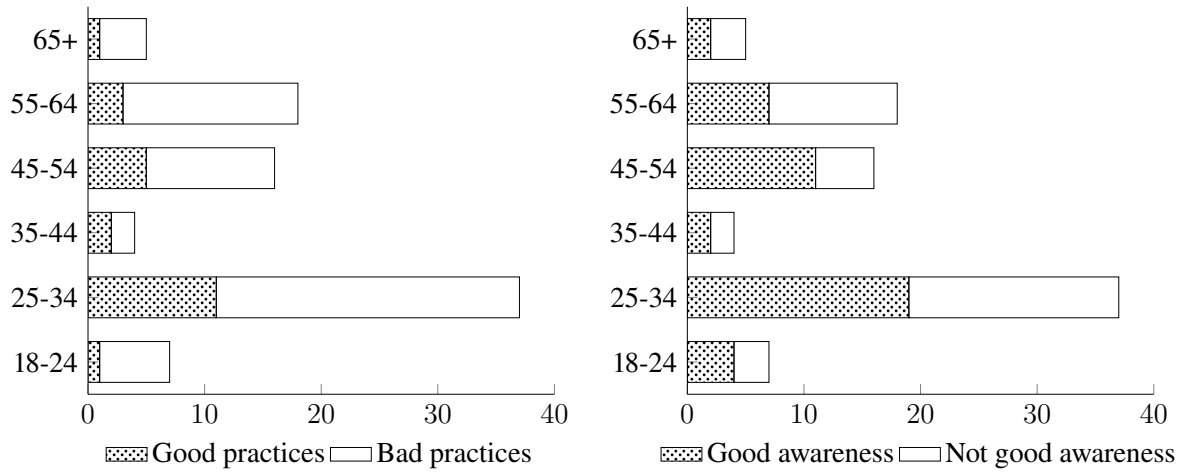


Figure 26. Cyber hygiene practices: systems updating and security awareness correlation with age.

Passwords, automatic locking are used by most people, see Figure 27. For part of Lithuanians, these are the right habits—turn off an unused item, or configured software for auto-locking or using a password. However, complex password generation is not in such a good situation; most people generate passwords only to meet the minimum requirements and use the same passwords on different websites. Most people lack an understanding of using strong passwords or even the importance of their data, which is not directly related to bank accounts. It could even be assumed that 68 % of those surveyed pose a threat to themselves and their loved ones, their companies. Due to inadequate password generation, as they give the attacker more straightforward access to the data.

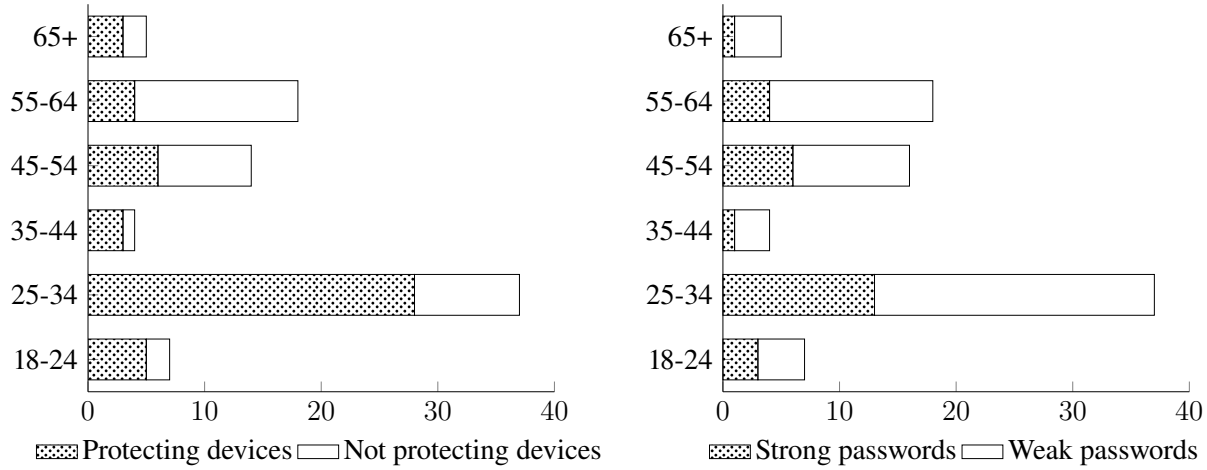


Figure 27. Cyber hygiene practices: protection of device and strong password correlation with age.

Many people in the study said they did not find themselves in situations where the money would be lured out of them. Many indicated that although an unknown person would introduce themselves to a high-ranking person or company employee from the IT department, they would not disclose the data to them. Distrust of strangers is probably based on the fact that people possibly work in not very large companies where they know each other. Also, after the general education of people in the press about fraud cases, distrust of unknown calls, emails these are grown in. As a result, as many as 91 % of respondents scored above 60 % (see Figure 28), which shows that at least so far, a person has not found himself in serious cyberattacks related to social engineering or is suspicious enough not to trust obscure information. While as many as 89 % of women and 94 % of men

are relatively resistant to at least the attacks they have heard of, most of them would disclose their personal information to their family members if there was a need. Therefore, loved ones are an essential part of cyber risk assessment. Also, both men and women do not have enough knowledge to generate strong passwords and promptly make systematic updates. These are unprotected areas that an attacker can effectively exploit.

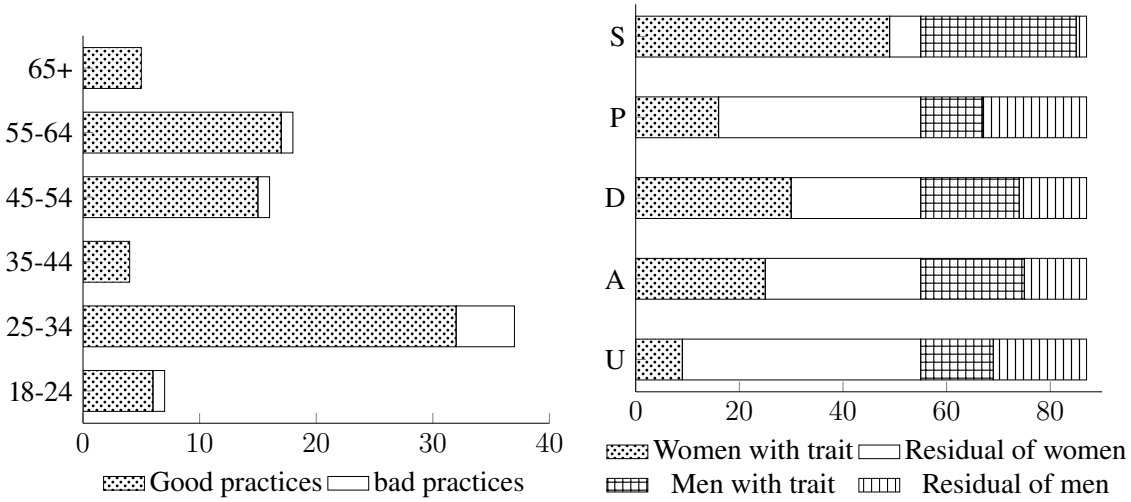


Figure 28. Resilience to social engineering attacks correlation with age in the left and cyber hygiene correlation with gender in right plot. U—system updating, A—security awareness, D—device protection, P—strong password generation, S—social security practices.

The article described 56 nations in the Big Five context [69], it has mentioned that Lithuanians are quite extroverted and open, and this study also showed such results. There are quite a bit of neurotic people among Lithuanians. However, this study's participants are also quite agreeable and conscientious because volunteering to participate in the survey is often agreed upon by pleasant enough, helpful people. And when assessing cyber risk, neurotics and open people have a higher risk than people with other characteristics based on privacy settings in social networks. Still, people with other factors are also at risk. For example, agreeableness makes them susceptible to manipulations with dark patterns. The survey did not show that there were many people with bright risk-taking preferences. Most people in this research area are with recreational and social risk-taking choices. Regarding situational decision-making questions, questionnaire results show that people make decisions depending on the situation, possibly determined by human values. Most respondents have a rational decision-making style.

4.2.2 Traits from photos

The question probably arises why age, gender, ethnicity, and other characteristics need to be detect from the photos, if there is a questionnaire. The aim is to gather as much information about the person as possible, without using a questionnaire and only visual, audio, and speech material on social networks. However, the data from which the algorithms can learn are needed first. Therefore, it is first necessary to gather a sufficient database of data on the human Big Five and other characteristics to be associated with facial features in the photos.

From the photos, it is possible to determine a lot of human characteristics, such as age, gender, ethnicity, specific traits, eyes, hair, skin color. There are many open-source projects available to recognize gender, age. However, the task to recognize age is not straightforward; humans may seem

different at the same age being from different regions. For example, the population of Greenland at the age of thirty may look like sixty year old due to malnutrition and possible genetic volatile. People who spend a lot of time in the sun also look older than others, and in old age, slender people look more wrinkled. Certain age groups are easier to identify, age groups from 0 to 13, and between 25 and 32 can be predicted with a higher reliability coefficient.

Although several algorithms are used to determine age, only 46 % of people are assigned to the appropriate age group, see Figure 29. Age is most correctly identified when people are younger. Approximately 78 % correctly identified the age that falls into the 25–34 age group. Also, people are most often assigned to a younger group than is actually the case, which can be explained by several reasons. First of all, the groups are relative, and if the age is a threshold, for example, if a person is 24 or 25—no one can distinguish it, so the error in one age group is not large. Secondly, people are most often assigned to a younger group. A possible variant is that the algorithms are trained to assign people to a younger group is less punishable than an older one. When a person is recognized as more youthful, he or she is less likely to take offense. For example, the study involved people over 65, and all of them were assigned to the age group between 55–64 (Figure 30(a)). Thirdly, such a situation can arise due to people's appearance, the environment in which people live—they can look younger or older, which already defines an individual nation's characteristics. It is possible to assume that Lithuanians look younger than other nations when comparing groups of the same age. The worst-established age, people, belonging to the 55–64 age group, is correctly identified in only one person, and some of them are even classified as youth groups. Most of these photos are of women, some more stocky, and therefore often look younger. Also, cosmetics can hide the signs of old age. There is also the fact that there is a difference in photo quality, lighting.

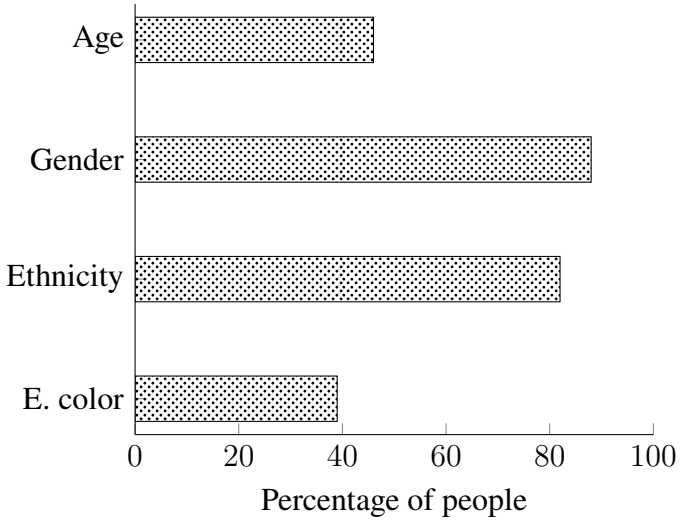


Figure 29. Percentage of correctly identified traits from pictures.

The human gender was best identified in the study, and even in 88 % of cases, the algorithms identified it correctly. All men were identified as men (see Figure 30(b)), with only a few women assigned to the male category. Except for one woman, all of them are older, most with short hair, without makeup. Perhaps this is because, in older age, the differences between people's sexual traits are narrowing. Ethnicity was correctly identified in about 82 % of cases. Latino-Hispanic and Asian are groups that have been assigned to the majority of misidentified people. Most smiling people whose eyes are narrowed in the photo because of the face mimicry was identified as Asian. Moreover, people who are more tanned, the face is rounder is attributed to Latino-Hispanic.



(a) The photo is an White woman older than 65, which is recognized as Latino-Hispanic woman age between 55–64.



(b) The photo is a White man with age between 25–34 years, which is recognized as a White man with age between 25–34.

Figure 30. Examples of successful and unsuccessful traits identification.

In the Figure 30, two examples of identification are shown, one successful and the other unsuccessful. Only the gender of a woman is correctly identified. As mentioned earlier, for all seniors over 65, the age group was found in a study between 55–64. The woman is assigned a Latino-Hispanic ethnicity. Algorithms set all the traits of the man correctly—age, gender, ethnicity. The gender of all men is correctly determined, and the age of younger people is correctly identified. According to the analyzed articles, gender determination is especially important, as most women belong to a higher cyber risk group. They are less resistant to phishing attacks and suffer from poor cyber hygiene. In the gender recognition aspect, the algorithms' overall accuracy is quite good, reaching 88 %. Young people (18–24, 25–34) at greater risk, as they spend much time with computer technology and often have little experience and knowledge of cybersecurity, so they do not have good security practices and are also quite vulnerable to phishing attacks. Algorithms quite accurately identify people belonging to youth groups.

4.2.3 Members of family recognition

An attacker can gather much information about the victim by examining photos on social networks. There are situations where the victim can be accessed through family members who are much more vulnerable, so to assess the full risk to the person, it is essential to assess which family members can be recognized as a kinship. The study involved 23 families and several individuals, and photos of the 22 families and individuals were analyzed. A total of 229 connections, of which 107 kinship, were obtained by examining family members within the family. The families also consist of a husband with a wife, but they are not related to blood ties, and in some places, they are uncles and nephews, nieces, but the algorithm was not trained to recognize this type of connection, this connection was marked as non-kinship.

Retrained FaceNet is used to determine the family relationship, with an accuracy of 75.6 % with test data. Unfortunately, such accuracy was not achieved with the study photos (accuracy—56 %). The result may be affected by the photos' quality, the angle of rotation of the face, and the lighting. Also, uncles, aunts, nieces, nephews, cousins, and other close relations were not taken as

blood connections for training. However, often and further relatives are very similar. The algorithm has identified not one such connection as a kinship—and this is true. Thus even several relations between the husband and wife were identified as blood. This occurrence can be based on the fact that often a person chooses the other human similar to himself. People living together become similar because they are affected by the same environment, eating the same food.

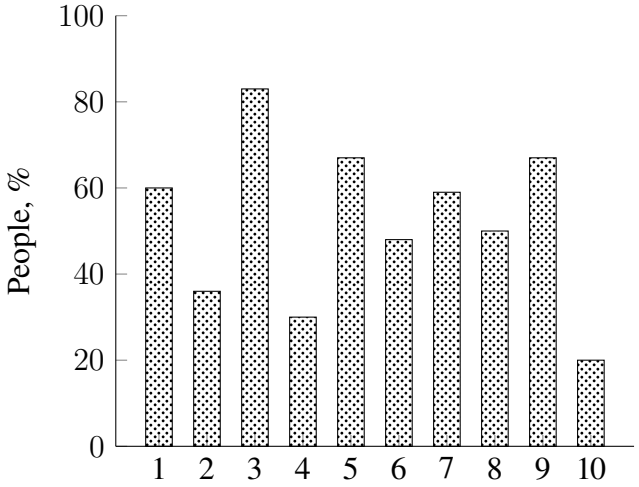


Figure 31. Different kinship recognition. 1—brothers, 2—siblings, 3—sisters, 4—father-daughter, 5—father-son, 6—mother-daughter, 7—mother-son, 8—grandfather-granddaughter, 9—grandmother-granddaughter, 10—grandmother-grandson.

In the study, the least relations were between grandparents and grandchildren, see Figure 31. There was no relation between grandfather and grandson. Therefore, the recognition results of these kinships are relative. The relationship between the mother and their children were at most, the accuracy of kinship recognition between mother and daughter was 48 %, and between mother and son was 59 %. Retrained neural network shows the best results in recognizing connections between sisters, achieving 83 % accuracy. This may also be related to the fact that the sisters' age in the study is quite similar.

This work is analyzed Lithuanian citizens—these are people living in the same climatic conditions, who have similar eating habits, so they are similar to each other. Comparing the population of different countries, the error of the trained algorithm would be less. Also, due to the similarity between husband and wife, it is related that the other half is chosen from the same region. And it is also observed that although Lithuania is a small country, regional differences can be seen in human physical traits, as it is based on lifestyle and common genetic features in that region. The identification of blood relations from photos is determined by many factors: the position of the face in the picture, enlightenment, age, gender, the similarity between people (defined by regional traits).

4.2.4 Traits correlation

This work examines how human characteristics are associated with cyber risk. And knowing the weaknesses, their combinations, a person can be presented with possible solutions in specific situations or just what qualities he needs to pay attention to. For example, athletes and how the Big Five traits related to their behavior in stressful situations were described, what methods helped them cope with the problems caused by the traits [65]. For example, for extroverted gymnasts, setting

a specific goal helps to maintain attention, and for those characterized by neuroticism, emotional control helps them cope with difficulties. Neuroticism is a trait that, in many situations, negatively affects a person's decisions and actions because people with this trait avoid decision-making, give in to stress, and do not cope with emotional tension. The article finds that, for example, neurotic cricket players proved more effective when they made sudden decisions. This assumption is possible because quick choices do not allow them to feel in doubt, to succumb to stress. And people with a sense of consciousness turn better in strategic situations when they focus on problems. Also, people who have a combination of characteristics: high extrovertability, openness, and low neuroticism cope with difficulties in a concentrated way. Thus, people with different traits sometimes have to behave differently in the same situations to achieve the desired result or avoid the problems.

Parallel coordinates diagram is designed to represent a multidimensional diagram in a 2D plot. This method of data representation was also applied in this research work to elucidate the possible correlations of certain features. Initially, an attempt was made to look at the general trends that can be made from traits identified in the questionnaire. Observed trends are shown in the plots. Because these are only trends, but it is not possible to distinguish a specific class and due to the abundance of data formed the so-called "Spaghetti chart". Therefore, to narrow the spectrum of data analysis, the analysis of family characteristics, an integral part of cybersecurity assessment, was chosen.

For better data comparison, numerical data was scaled. Since most of the study, people were identified as extroverted, to find trends were found median of extrovertability scores, and based on it, people were divided into more extroverted and less extroverted. In Figure 32 depicts traits, cyber hygiene habits with which specific trends were observed in traits correlation. Extroverted people

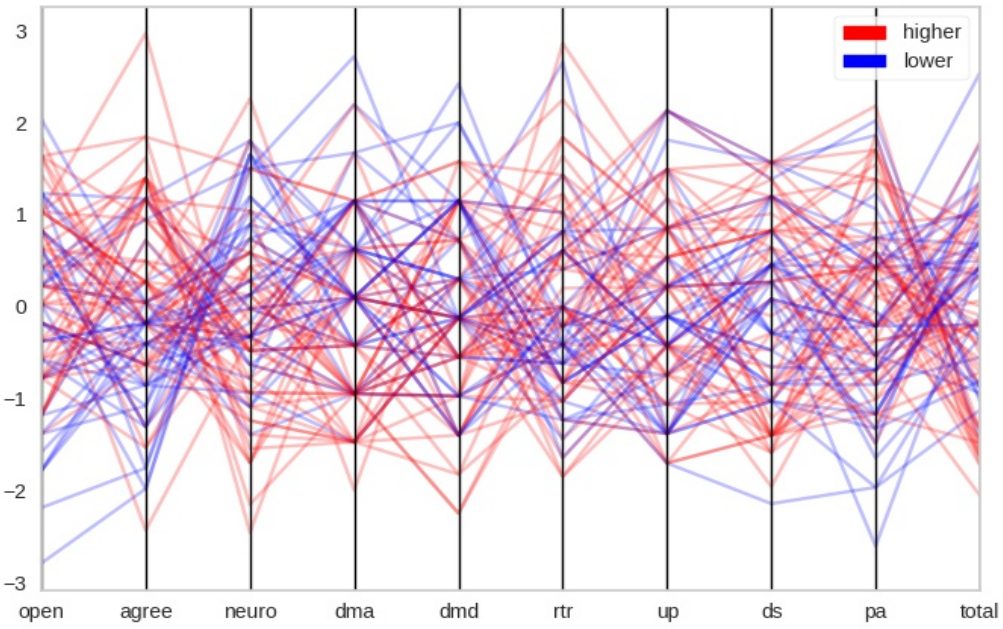


Figure 32. Extroversion correlation with other traits. The vertical axis contains scaled score points. Abbreviations: open—openess, agree—agreeableness, neuro—neuroticism, dma—avoidant and dmd—dependant decision making styles, rtr—recreational risk-taking preference, up—updating, ds—device securement, pa—protective awareness, total—risk score.

are more likely to score high openness, agreeability scores; they are less neurotic. Extrovert more often does not avoid making decisions, and they do not need other opinions to make them. They also love more extreme entertainment and are not afraid to take risks in this area. More introverted people are more likely to report poorer updating and security awareness practices but better device securement. Extroverted ones show good enough cyber hygiene. But high extrovertability with low neuroticism and high agreeableness, openness scores increase social engineering risk. Neurotics, due to their suspicion, distrust, are less susceptible to social engineering attacks. However, all other Big Five traits under The "Social Engineering Personality Framework" (SEPF) [81] increase the risk by increasing the potential number of principles, based on them, can be created social engineering scenarios. Therefore, an attacker can access a person with the above characteristics based on all six principles: liking, authority, reciprocity, social proof, scarcity, commitment, and consistency.

Based on questions that had to identify people's potential susceptibility to social engineering attacks, no association with Big Five traits was observed. However, trends related to the decision-making style were observed (see Figure 33). People who think they would be resistant to social

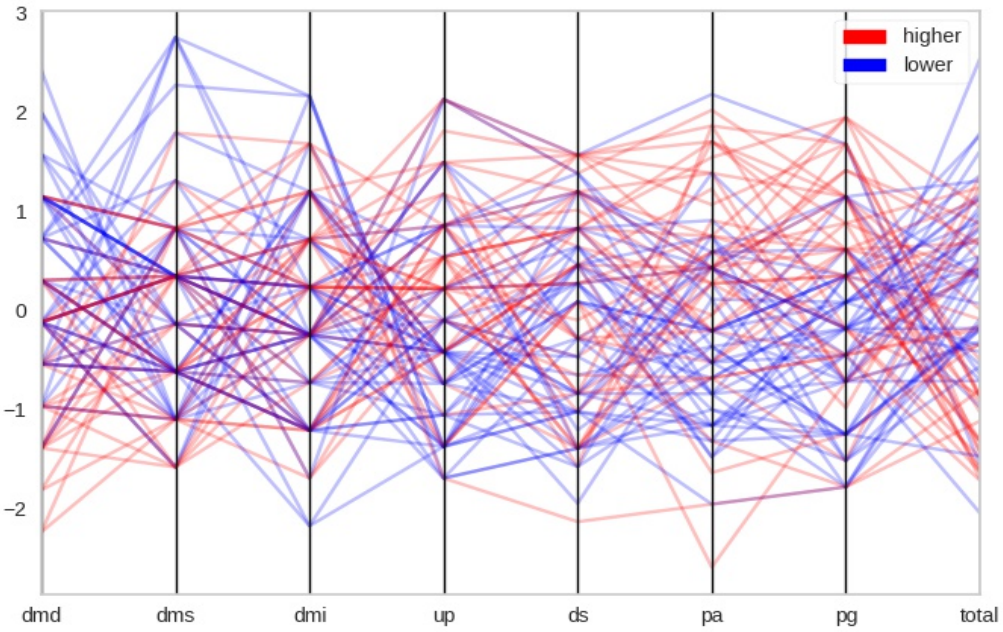


Figure 33. Resilience to social engineering attacks correlation to other traits. The vertical axis contains scaled score points. Abbreviations: dmd—dependant, dms—spontaneous and dmi—intuitive decision making styles, rtr—recreational risk-taking preference, up—updating, ds—device securement, pa—protective awareness, pg—password generation, total—risk score.

engineering attacks make less spontaneous or less dependent on others' opinions decisions. There is also a tendency to rely on intuition. Of course, these people are potentially more resistant to social engineering attacks because they may have more cybersecurity knowledge. The parallel coordinates plot shows that their cyber hygiene habits are good enough.

In analyzing cyber hygiene practices, trends have been observed with several traits. People with lower security awareness scores are more neurotic (see Figure 34). And those with higher scores of security awareness make more rational decisions, less avoidant, spontaneous, or depen-

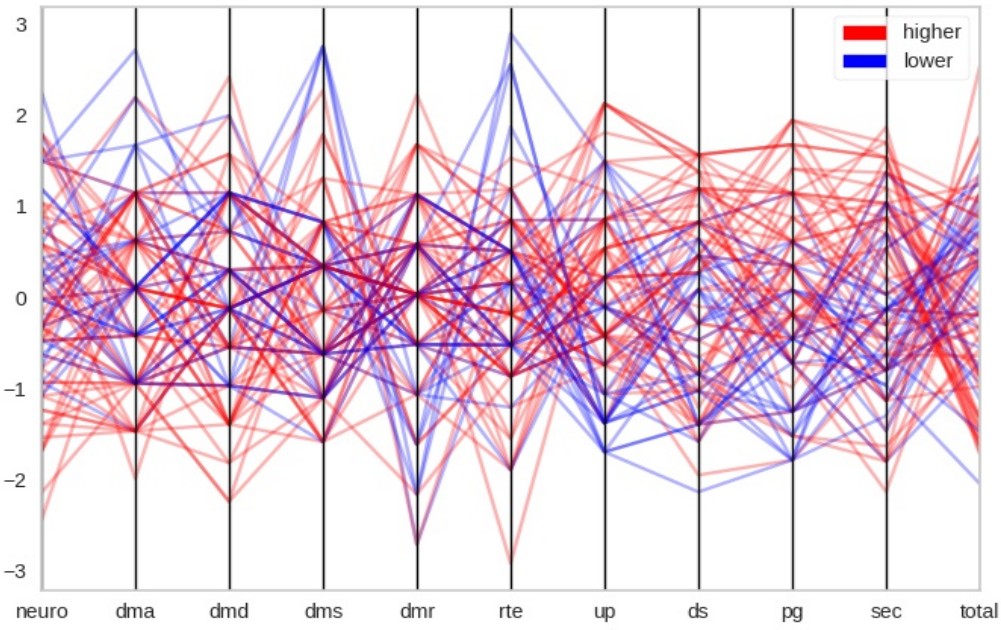


Figure 34. Protective awareness correlation with other traits. The vertical axis contains scaled score points. Abbreviations: neuro—neuroticism, dma—avoidant, dmd—dependant, dms—spontaneous, dmr—rational decision making styles, rte—ethical risk-taking preference, up—updating, ds—device securement, pg—password generation, sec—other security practices, total—risk score.

dant decisions. They also have a low ethical risk-taking preference score. Also, from the parallel coordinates plots examined, it is assumed that most often, if a person adheres to cyber hygiene in one place, his/her update and device security and other practices are adequate. For example, people with a high-security awareness score also generate strong passwords and perform periodic updates quickly. Such traits complex: good cyber hygiene practices, unhurried, thoughtful decision-making, neuroticism is associated with low cyber risk, low probability of social engineering attack.

A high openness rate is also associated with better cyber hygiene practices, less susceptibility to social engineering attacks. Correct update practices, device protection, high-security awareness scores are positively correlated with strong password generation (see Figure 35). Which also pos-

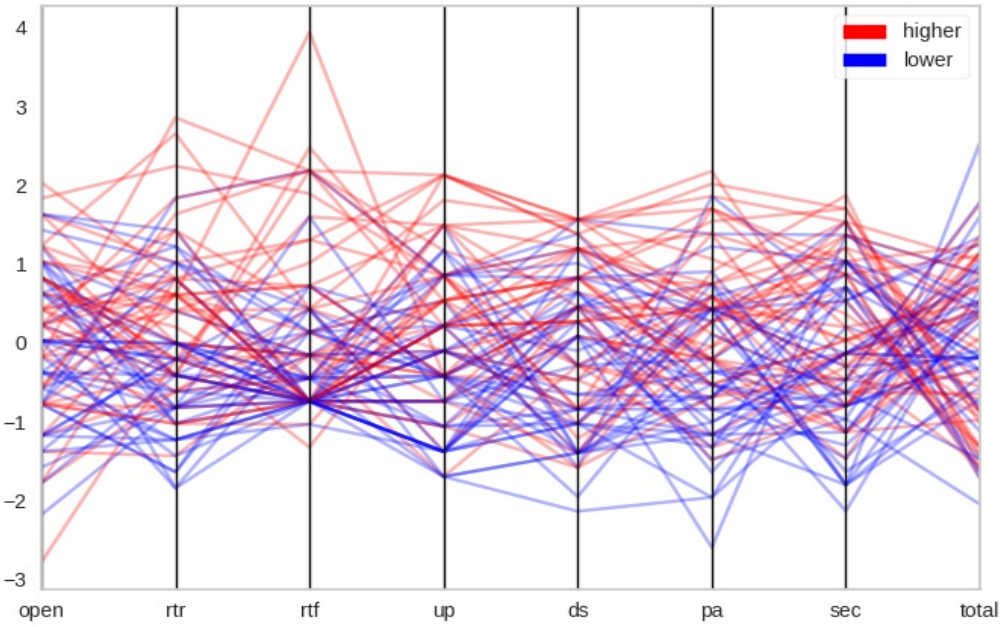


Figure 35. Strong password generation correlation with other traits. The vertical axis contains scaled score points. Abbreviations: open—openness, rtr—recreational, rtf—financial risk-taking preferences, up—updating, ds—device securement, pa—protective awareness, sec—other security practices, total—risk score.

itively correlates with openness. It is related to the fact that open people are open to innovation; they have a more comprehensive knowledge range. But there are also threats—they do not always follow security policies, their privacy settings are low.

A couple of families were randomly selected, assessing the correlation of family traits. Their scores were not scaled to show that the scores collected among family members are quite similar. Traits score scales vary; Big Five has a maximum score of 50, decision-making style 15, risk-taking preferences 30, updating 15, device securement 25, security awareness 45, password generation 20, resilience to social engineering attack 70. Children's scores are identical in some cases to mother, others to father results. Traits are similar between family members, such as Big Five due to genetic heredity, environment influences, value education. For example, family members No. 15 are less extroverted and more closed than most family members No. 6, and most members of this family are more conscientious and less neurotic compared to the other two families (see Figure 36). It

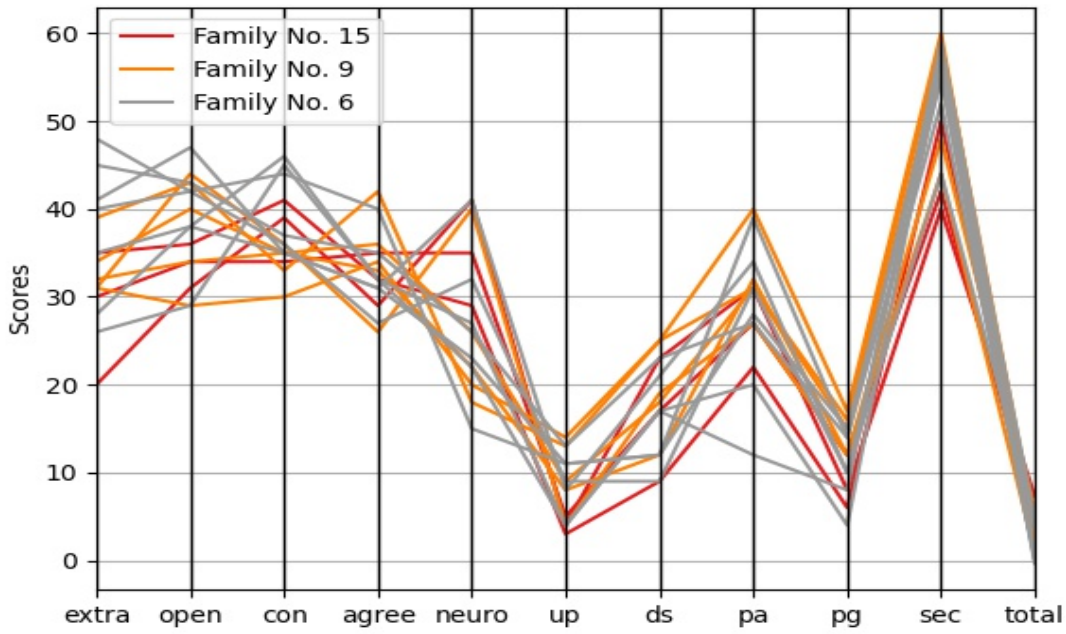


Figure 36. Comparison of three families in Big Five traits and cyber hygiene context. Abbreviations: extra—extrovertability, open—openness, con—conscientiousness, agree—agreeableness, neuro—neuroticism, up—updating, ds—device securement, pa—protective awareness, pg—password generation, sec—other security practices, total—risk score.

also appears from the chart data that extrovertability is negatively correlated in many people with neuroticism. Although security practices are an acquired trait, similarities are observed between family members in this area. The majority of Family No. 9 have better cyber hygiene practices than members of the other two families. This situation is most likely related to the knowledge transmitted.

The study was involved three members of Family No. 15 who are related by close kinship: parents—children. It can be a reason for the similarity of the traits of the members of Family No. 15 is especially noticeable. As well as two of the members of the family are twins. Members of this family make a more avoidant but less spontaneous or intuition-based decision (see Figure 37). And it differs from other families in that they have higher dependent decision-making style scores. What shows the close relationship between family members, the dependence of each other. Therefore, family members' cyber risk is possibly higher, as family members can be accessed by accessing one family member. In other families, the scatter is on a larger amplitude. Like Family No. 6 is making less avoidant, spontaneous, dependent, but more rational or intuition-based decisions. Thus, the similarity between the members of the family is observed. In one aspect, it is brighter, in others as risk-taking preferences are less noticeable.

A more detailed distribution of family No. 4 member scores in the parallel coordinates diagram can be seen in Figure 38. The resulting curves are very similar. It can also be seen that the parental curves are quite similar—this, like the results of kinship identification, shows that similar people create a family. And the results of children's traits are, in some cases, utterly identical to one of the parents' outcomes. The Big Five features are characteristic of a person if he scores more than 60 %, so it can be seen that the No. 4 family members are extroverted, agreeable, and quite conscientious.

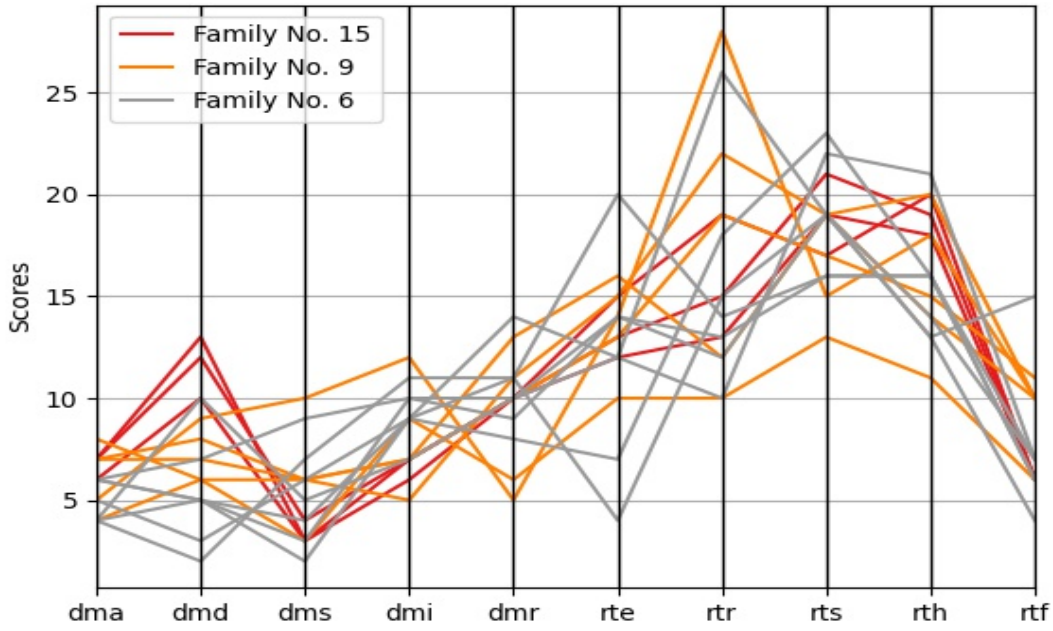


Figure 37. Comparison of three families in decision making style and risk-taking preferences context. Abbreviations: dma—avoidant, dmd—dependant, dms—spontaneous, dmi—intuitive and dmr—rational decision making styles, rte—ethical, rtr—recreational, rts—social, rth—health/safety and rtf—financial risk taking preferences.

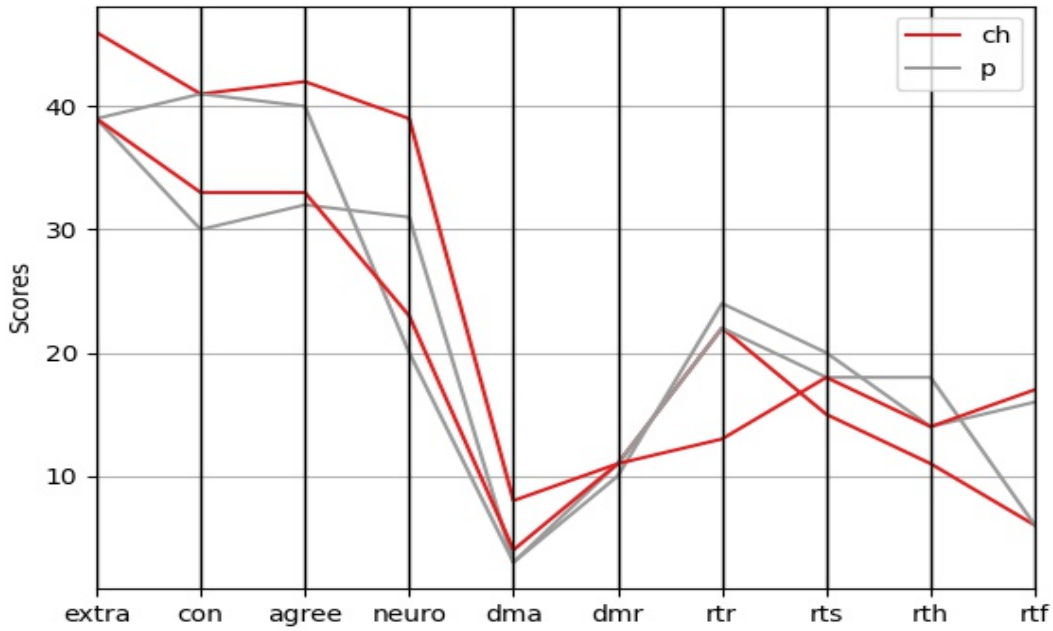


Figure 38. Comparison of traits of parents (p) and children (ch) of one family. The vertical axis contains score points. Abbreviations: extra—extrovertability, con—conscientiousness, agree—agreeableness, neuro—neuroticism, dma—avoidant and dmr—rational decision making styles, rtr—recreational, rts—social, rth—health/safety and rtf—financial risk taking preferences.

One child and father are more neurotic than the other two family members. All family members make rational decisions at a similar level; most of them like extreme entertainment, and two family members are inclined to take financial risks, the other two are not. An attacker knowing one member of the family can assume that other family members also love the same things or whether they will behave similarly in similar situations.

Some trends were observed, such that extroverts are more open to innovation, more agreeable, and more emotionally stable, analyzing the data. People who follow acceptable cyber hygiene practices in one area often act responsibly in others. Members of the family have similar traits and are inclined to make similar decisions, which can be exploited during social engineering attacks. However, certain combinations of characteristics can protect people from cyberattacks: for example, increased neuroticism, propensity for conscientious, rational decision-making style, good cyber hygiene practices. However, more personalized training based on identified characteristics could reduce each person's cyber risk.

Results and Conclusions

There are many machine training algorithms, neural networks that can recognize various human features from visual, aural, and written data. This work describes an ontology-based model, a semantic network that structures existing data. The summarization of the data simplifies the possible implementation of further research. In this work, based on a semantic network, a cybersecurity risk assessment prototype was developed, which can be used to assess the risk by evaluating a persons' characteristics from answers to the questions. From photographs we are identifying such traits as age, gender, or family members. Traits evaluation coefficients were calculated based on the analyzed literature. This human risk assessment prototype is a preliminary version that could be extended to include more and more traits recognition. Already collecting data, the dataset is created for other research, such as the possibility of associating the traits of human character with the details of appearance.

This study also conducted a risk assessment study of 22 families. The study differs from other works that involved not only members of the university community but also families. Most of whom members have no education in IT technology and little knowledge of cybersecurity. From the survey results, it is possible to make assumptions characterizing of a certain part of the population. Lithuanians who voluntarily agree to participate in surveys are more extroverted, open, sufficiently agreeable, conscientious, and are a significant part of neurotics. A minority of people tend to take risks; Lithuanians take little risks even in recreational or social areas. Also, most people do not have a bold decision-making style, which indicates that decisions are made based on the situation. It was most evident that people were more likely to make rational decisions. Such a common set of population traits indicates that Lithuanians should be quite resistant to cyberattacks based on the researched aspects. However, according to survey data on cyber habits, most people lack cybersecurity knowledge, even young people who spend a lot of time with computer technology. Cybersecurity training should be included in the early learning process, that is, in school curricula.

Trends in the correlation between traits have been observed, but they cannot be described as rules. However, a strong resemblance of character traits is observed among kinships. Children take characteristics from one parent; this is related to genetics and the upbringing principles created by the environmental conditions. Therefore, family members' identification is a rather important aspect in assessing human risk, as family members pose a dual-threat. They can directly provide information about the victim of a cyberattack, and the behavior of family members can be assumed how the victim will behave in similar situations. Several neural networks were trained for kinship identification, but any of the best results were not as good as we would like them to be. There is an aspect to comparing members of a joint family that people often come from the same regional location, which means that these people have more similarities for that reason alone. Although some assumptions have been made about kinship recognition results, it would still be useful to test the models trained in other projects and see how they deal with recognizing people who are only from certain regions.

Many different factors cause human cyber risk. Some of them are evaluated in this work. Still, the evaluation can be developed in several aspects—sources from which the traits are determined, elements according to which the risk scores are given, a spectrum of the evaluated features. The assessment of individual human disorders alone could provide an even more accurate assessment of the level of risk in the context of genetics. A broader evaluation of genetic traits would potentially reveal hidden features that unfold in certain situations. This would help determine if the person

will behave in the situation as he described in the question-answer or it is just his opinion. Human risk assessment from traits is a broad topic, so the results obtained are only primary data for further research.

Future Works

The assessment of digital traits in cybersecurity is vast, so this work should be developed in several aspects. The spectrums of recognizable features and risk scenarios should be widening. Machine training methods could identify human characteristics from posts on social networks, audio recordings, and facial proportions. More hidden features related to genetics can be identified, and traits should be associated with genes. At the genetic level, one characteristic is often associated with another, thus making it possible to create a more accurate set of human traits that would affect cyberattacks' reaction. Possible protection methods or attack scenarios presented should be after risk assessment. Also, to compile a data set for further research, which would be suitable not only for the people of the Lithuanian region, but it would also be necessary to assess the risk of other countries' population. Also, a clearer picture of the correlation of traits would be obtained by analyzing the employees of different companies so that the participants would have more varied characteristics of the character.

Kinship recognition is important in cyber risk assessment, so other neural networks should be tested in the context of this task, or even more, variation should be tested in training the current solution. Kinship-related people have similar features that are determined by genetics, so if necessary, it would be possible to examine the cyber risks posed by genetics in the context of families. However, there would be a massive amount of research work here. Many of the potentially significant characteristics of a person would be revealed. Identifying family members and that they may increase the risk may be sufficient for the cyber risk assessment. A possibly better option would be to develop risk scenarios based on the traits' results to draw people's attention to their weaknesses.

References

[1] Olusola Akinrolabu, Jason R.C. Nurse, Andrew Martin, and Steve New. Cyber risk assessment in cloud provider environments: Current models and future needs. *Computers & Security*, 87:101600, 2019.

[2] Jhan S Alarifi, Manu Goyal, Adrian K Davison, Darren Dancey, Rabia Khan, and Moi Hoon Yap. Facial skin classification using convolutional neural networks. In *International Conference Image Analysis and Recognition*, pages 479–485, 07 2017.

[3] Rami Suleiman Alkhawaldeh. Dgr: Gender recognition of human speech using one-dimensional conventional neural network. *Scientific Programming*, 2019:1–12, 2019.

[4] Felix Anda, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. Evaluating automated facial age estimation techniques for digital forensics. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 129–139, 05 2018.

[5] Baigio Antonelli. Ethnicity classifier. https://github.com/BiagioAntonelli/Ethnicity_classifier. Last seen: 2020-06-08.

[6] Ann-Renee Blais and Elke Weber. A domain-specific risk-taking (dospert) scale for adult populations. *Judgment and Decision Making*, 1:33–47, 02 2006.

[7] Diana Borza, Razvan Itu, Radu Danescu, and Ioana Barbantan. *Analysing Facial Features Using CNNs and Computer Vision*, pages 146–157. Springer, 01 2020.

[8] Thirimachos Bourlai, Ayman Abaza, Arun Ross, and Mary Harrison. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biometrics*, 3:314–324, 12 2014.

[9] Ashley A. Cain, Morgan E. Edwards, and Jeremiah D. Still. An exploratory study of cyber hygiene behaviors and knowledge. *Journal of Information Security and Applications*, 42:36–45, 2018.

[10] P. F. Carrera. Face recognition algorithms, 2010. http://www.ehu.eus/ccwintco/uploads/e/eb/PFC-IonMarques.pdf. Last seen: 2020-06-16.

[11] National Cybersecyrity centre with the Ministry of Defence of Lithuania. 2018 the report of the cyber security state, 2018. https://www.nksc.lt/doc/NKSC_ataskaita_2018.pdf. Last seen: 2020-03-29.

[12] Kang Leng Chiew, Kelvin Sheng Chek Yong, and Choon Lin Tan. A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, 106:1–20, 2018.

[13] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 07 2017.

[14] Mark Ciampa. *CompTIA Security+ Guide to Network Security Fundamentals, Loose-Leaf Version*. Course Technology Press, Boston, MA, USA, 5th edition, 2015.

[15] Casey Cole, Bethany Janos, Dien Anshari, James Thrasher, Scott Strayer, and Homayoun Valafar. Recognition of smoking gesture using smart watch technology. In *Health Informatics and Medical Systems (The 2016 WorldComp International Conference Proceedings)*, 07 2016.

[16] Teresa Correa, Joe Straubhaar, Wenhong Chen, and Jeremiah Spence. Brokering new technologies: The role of children in their parents' usage of the internet. *New Media & Society*, 17:483–500, 03 2013.

[17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[18] Adrian Sergiu Darabant Diana Borza and Radu Danescu1. Real-time detection and measurement of eye features from color images. *Sensors*, 16:1105, 07 2016.

[19] Sopagna Eap, D. Degarmo, Ayaka Kawakami, Shelley N Hara, G. C. Hall, and A. Teten. Culture and personality among european american and asian american men. *Journal of Cross-Cultural Psychology*, 39:630–643, 2008.

[20] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2873—-2882, New York, NY, USA, 2015. Association for Computing Machinery.

[21] Hannah Foldes, EMILY DUEHR, and Deniz Ones. Group differences in personality: Meta-analyses comparing five u.s. racial groups. *Personnel Psychology*, 61:579–616, 09 2008.

[22] Félix Fuentes, Jose Diego-Mas, Valery Naranjo, and Mariano Alcañiz Raya. Automatic classification of human facial features based on their appearance. *PLOS ONE*, 14:e0211314, 01 2019.

[23] Brandon Gavett, Rui Zhao, Samantha John, Cara Bussell, Jennifer Roberts, and Chuan Yue. Phishing suspiciousness in older and younger adults: The role of executive functioning. *PLOS ONE*, 12:e0171620, 02 2017.

[24] Margaret Gratian, Sruthi Bandi, Michel Cukier, Josiah Dykstra, and Amy Ginther. Correlating human traits and cybersecurity behavior intentions. *Computers & Security*, 73:345—-358, 11 2017.

[25] P. Grother, M. Ngan, K. Hanaoka, Information Technology Laboratory (National Institute of Standards, and Technology). *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification*. NISTIR; NIST IR; NIST interagency report; NIST internal report. U.S. Department of Commerce, National Institute of Standards and Technology, 2018.

[26] Vikas Gupta. Face detection – opencv, dlib and deep learning (c++ / python), 2018. https://www.learnopencv.com/face-detection-opencv-dlib-and-deep-learning-c-python/. Last seen: 2020-06-16.

[27] Mehmet Hacibeyoglu and Mohammed Hussein Ibrahim. Human gender prediction on facial mobil images using convolutional neural networks. *International Journal of Intelligent Systems and Applications in Engineering*, 6(3):203–208, 09 2018.

[28] Tzipora Halevi, James Lewis, and Nasir Memon. A pilot study of cyber security and privacy related behavior and personality traits. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 737—744, New York, NY, USA, 2013. Association for Computing Machinery.

[29] Darren R. Hayes and Francesco Cappa. Open-source intelligence for risk assessment. *Business Horizons*, 61(5):689 – 697, 2018.

[30] Robin Hennessy, Patrizia Baldwin, David Browne, Anthony Kinsella, and John Waddington. Frontonasal dysmorphology in bipolar disorder by 3d laser surface imaging and geometric morphometrics: Comparisons with schizophrenia. *Schizophrenia research*, 122:63–71, 09 2010.

[31] Diane Henshel, M.G. Cains, Blaine Hoffman, and Timothy Kelley. Trust as a human factor in holistic cyber security risk assessment. *Procedia Manufacturing*, 3:1117–1124, 12 2015.

[32] Romuald Hoffmann. Stochastic model of the simple cyber kill chain: Cyber attack process as a regenerative process. In Khalid Saeed and Jiří Dvorský, editors, *Computer Information Systems and Industrial Management*, pages 355–365, Cham, 2020. Springer International Publishing.

[33] Eric M Hutchins, Michael J Cloppert, Rohan M Amin, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, 1(1):80, 2011.

[34] International Personality Item Pool. Administering ipip measures, with a 50-item sample questionnaire, 2019. https://ipip.ori.org/New_IPIP-50-item-scale.htm. Last seen: 2020-10-25.

[35] Ankita Jain and Vivek Kanhangad. Gender classification in smartphones using gait information. *Expert Systems with Applications*, 93:257–266, 10 2017.

[36] Zhifeng Li Kaipeng Zhang, Zhanpeng Zhang and Yu Qiao. Joint face detection and alignment using mult-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[37] Jessica L. Kamerer and Donna McDermott. Cybersecurity: Nurses on the front line of prevention and education. *Journal of Nursing Regulation*, 10(4):48 – 53, 2020.

[38] Christian Kandler, Wiebke Bleidorn, Rainer Riemann, Alois Angleitner, and Frank Spinath. The genetic links between the big five personality traits and general interest domains. *Personality and Social Psychology Bulletin*, 37:1633–1643, 11 2011.

[39] keras.io. Keras applications. https://keras.io/api/applications/. Last seen: 2020-10-25.

[40] Murat Kezer, Barış Sevi, Zeynep Cemalcilar, and Lemi Baruh. Age differences in privacy attitudes, literacy and privacy management on facebook. *Cyberpsychology*, 10, 05 2016.

[41] Vladimir V. Khryashchev and Anatoly Sedov. Development of face image quality assessment algorithms for biometric identification tasks, 2017. https://csit.am/2017/Proceedings/PRIP/PRIP4.pdf. Last seen: 2020-12-09.

[42] Inseong Kim, Joon Hyung Shim, and Jinkyu Yang. Face detection. https://web.stanford.edu/class/ee368/Project_03/Project/reports/ee368group02.pdf. Last seen: 2020-06-16.

[43] Gil Levi and Tal Hassncer. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 06 2015.

[44] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[45] Ponemon Institute LLC. Ninth annual cost of cybercrime study unlocking the value of improved cybersecurity protection, 2019. https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf. Last seen: 2020-01-25.

[46] Zhipeng Luo, Zhiguang Zhang, Zhenyu Xu, and Che Lixuan. Challenge report: Recognizing families in the wild data challenge, 05 2020. http://128.84.4.18/pdf/2006.00154. Last seen: 2020-10-25.

[47] Mateus Mendes Maksat Kanatov, Lyazzat Atymtayeva. Improved facial expression recognition with xception deep net and preprocessed images. *Applied Mathematics & Information Sciences*, 13:859–865, 09 2019.

[48] Satya Mallick. Code for age gender recognition using deep learning. https://github.com/spmallick/learnopencv/tree/master/AgeGender. Last seen: 2020-04-19.

[49] Joyce A. Martin, Brady E. Hamilton, Michelle J K Osterman, Anne K. Driscoll, and Patrick Drake. Births: Final data for 2018. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 68 13:1–47, 2019.

[50] Agata McCormac, Tara Zwaans, Kathryn Parsons, Dragana Calic, Marcus Butavicius, and Malcolm Pattinson. Individual differences and information security awareness. *Computers in Human Behavior*, 69:151–156, 12 2016.

[51] Afef Mdhaffar, Fedi Cherif, Yousri Kessentini, Manel Maalej, Jihen Ben Thabet, Mohamed Maalej, Mohamed Jmaiel, and Bernd Freisleben. Dl4ded: Deep learning for depressive episode detection on mobile devices. In José Pagán, Mounir Mokhtari, Hamdi Aloulou, Bessam Abdulrazak, and María Fernanda Cabrera, editors, *How AI Impacts Urban Living and Public Health*, pages 109–121, Cham, 2019. Springer International Publishing.

[52] Mats Nagel, Philip Jansen, S Stringer, K Watanabe, Christiaan Leeuw, Julien Bryois, Jeanne Savage, Anke Hammerschlag, Nathan Skene, AB Muñoz-Manchado, M Agee, B Alipanahi, A Auton, RK Bell, K Bryc, SL Elson, P Fontanillas, NA Furlotte, DA Hinds, and D Posthuma. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics*, 50:920—-927, 07 2018.

[53] F Navabifar, M. Emadi, R. Yusof, and M. Khalid. What are the most common cyber attacks? https://www.cisco.com/c/en/us/products/security/common-cyberattacks.html. Last seen: 2020-06-16.

[54] Farhad Navabifar, Mehran Emadi, Rubiyah Yusof, and Marzuki Khalid. A short review paper on face detection using machine learning, 2010. https://pdfs.semanticscholar.org/114c/8ceba9ded2a7f733d0f3ba64f380b2bb7786.pdf. Watch: 2020-06-16.

[55] Northeastern SMILE Lab. Northeastern smile lab - recognizing faces in the wild, 2019. https://www.kaggle.com/c/recognizing-faces-in-the-wild/leaderboard. Last seen: 2020-10-25.

[56] Federal Bureau of Investigation Internet Crime Complaint Center (ICE3). 2018 internet crime report, 2019. https://pdf.ic3.gov/2018_IC3Report.pdf. Last seen: 2020-01-25.

[57] Opencv 2.4.13.7 documentation. https://docs.opencv.org/2.4/modules/objdetect/doc/cascade_classification.html. Last seen: 2020-06-16.

[58] Histogram of oriented gradients. https://www.learnopencv.com/histogram-of-oriented-gradients/. Last seen: 2020-06-16.

[59] Yuval Palgi, Amit Shrira, Menachem Ben-Ezra, Sara Fridel, and Ehud Bodner. The relationships between daily optimism, daily pessimism, and affect differ in young and old age. *Personality and Individual Differences*, 50:1294–1299, 06 2011.

[60] Malcolm Pattinson, Cate Jerram, Kathryn Parsons, Agata McCormac, and Marcus Butavicius. Why do some people manage phishing e-mails better than others? *Information Management & Computer Security*, 20:18–28, 03 2012.

[61] Eiluned Pearce, Rafael Wlodarski, Anna Machin, and Robin Dunbar. Genetic influences on social relationships: Sex differences in the mediating role of personality and social cognition. *Adaptive Human Behavior and Physiology*, pages 331—351, 11 2019.

[62] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of sparql. In *International semantic web conference*, pages 30–43. Springer, 2006.

[63] Robert Power and Michael Pluess. Heritability estimates of the big five personality traits based on common genetic variants. *Translational psychiatry*, 5:e604, 07 2015.

[64] Dan Raywood. Bhusa dropped usb experiment detailed, 2016. https://www.infosecurity-magazine.com/blogs/bhusa-dropped-usb-experiement/. Last seen: 2020-03-28.

[65] Ross Roberts and Tim Woodman. Personality and performance: Moving beyond the big 5. *Current Opinion in Psychology*, 16:104–108, 05 2017.

[66] Joseph Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:2624–637, 04 2018.

[67] Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F. Xavier Roca, and Jordi Gonzàlez. Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 72:563 – 571, 2017.

[68] Koustuv Saha and Munmun Choudhury. Modeling stress with social media around incidents of gun violence on college campuses. *Proceedings of the ACM on Human-Computer Interaction*, 1:92:1–, 11 2017.

[69] David Schmitt, Jüri Allik, Robert McCrae, Veronica Benet, Lidia Alcalay, Lara Ault, Ivars Austers, Kevin Bennett, Gabriel Bianchi, Fredrick Boholst, Mary Cunen, Johan Braeckman, Edwin Jr, Leo Gerard Caral, Gabrielle Caron, Maria Casullo, Michael Cunningham, Ikuo Daibo, Charlotte de backer, and Agata Zupanèiè. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, 38:173–212, 01 2007.

[70] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 06 2015.

[71] Scikit-learn. support vector machines, 2019. https://scikit-learn.org/stable/modules/svm.html. Last seen: 2020-06-16.

[72] Susanne Scott and Reginald Bruce. Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement - EDUC PSYCHOL MEAS*, 55:818–831, 10 1995.

[73] Sefik Ilkin Serengil. Deepface. https://github.com/serengil/deepface. Last seen: 2020-04-19.

[74] Sandeep Sharma Shaily Pandey. Review: Face detection and recognition techniques. *International Journal of Computer Science and Information Technologies*, 5(3):4111–4117, 2014.

[75] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 373—-382, New York, NY, USA, 2010. Association for Computing Machinery.

[76] Kyuyong Shin, Kyoung Min Kim, and Jongkwan Lee. A study on the concept of social engineering cyber kill chain for social engineering based cyber operations. *Journal of The Korea Institute of Information Security & Cryptology*, 28(5):1247–1258, 2018.

[77] Jordan Shropshire, Merrill Warkentin, Allen Johnston, and Mark Schmidt. Personality and it security: An application of the five-factor model. In *AMCIS*, volume 6, page 415, 01 2006.

[78] Olav B Smeland, Yunpeng Wang, Min-Tzu Lo, Wen Li, Oleksandr Frei, Aree Witoelar, Martin Tesli, David A Hinds, Joyce Y Tung, Srdjan Djurovic, et al. Identification of genetic loci shared between schizophrenia and the bigfive personality traits. *Scientific reports*, 7:2222, 2017.

[79] Toyoland, 2008. https://www.toyoland.com/toyota/precrash.html. Last seen: 2020-06-16.

[80] Yusuke Uchida. Age and gender estimation. https://github.com/yu4u/age-gender-estimation. Last seen: 2020-04-19.

[81] Sven Uebelacker and Susanne Quiel. The social engineering personality framework. In *2014 Workshop on Socio-Technical Aspects in Security and Trust*, pages 24–30, 07 2014.

[82] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137—154, 5 2004.

[83] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114:246–257, 02 2018.

[84] Gary Weiss, Jessica Timko, Catherine Gallagher, Kenichi Yoneda, and Andrew Schreiber. Smartwatch-based activity recognition: A machine learning approach. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 426–429, 02 2016.

[85] Jin You, Helene Fung, and Derek Isaacowitz. Age differences in dispositional optimism: A cross-cultural study. *European Journal of Ageing*, 6:247–252, 12 2009.

Appendices

A Semantic network

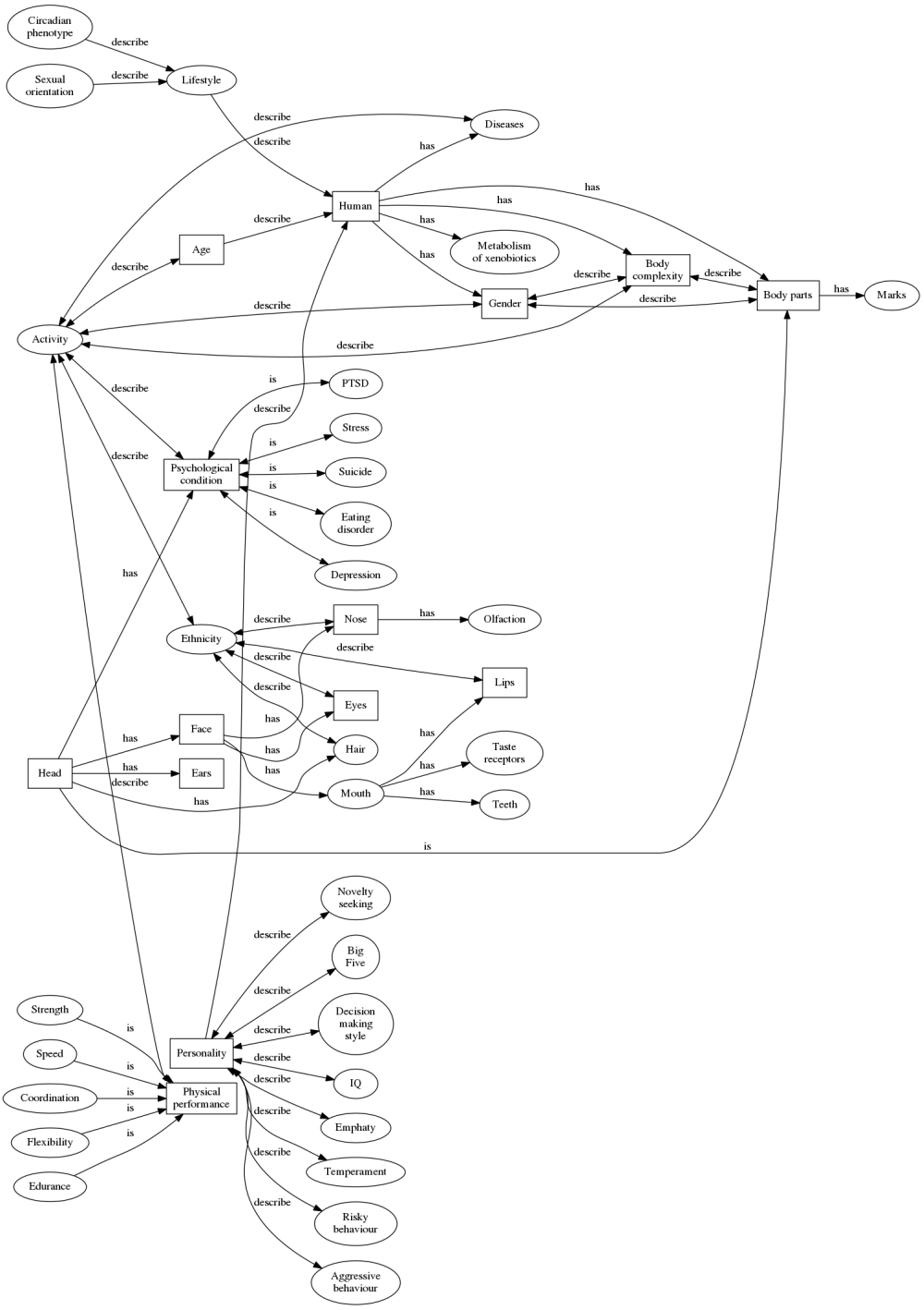


Figure 39. Semantic network of traits, algorithms and risk values are omitted