Sofia University “St. Kliment Ohridski”
Faculty of Mathematics and Informatics

Intelligent information systems in bioinformatics: semantic integration, analysis and classification of biomedical data

Ilian Nedkov Mihailov

Thesis summary

for awarding with the educational and scientific degree "PhD "
in professional domain 4.6 Informatics and Computer Science ”,
PhD program “Information Technologies - Bio and Medical Informatics ”

Supervisor: Assoc. prof. Dimitar Ivanov Vassilev, PhD.

Sofia, 2021.

Where not indicated, all references in the text of the abstract to pages, chapters, sections, cited literature, tables and figures refer to the text of the thesis. All figures, tables and cited literature in the abstract use the same numbering as in the text of the thesis.

Contents

General characteristics of the thesis

Content, objective and structure of the thesis The

The thesis is written on 186 pages, which include 50 figures, 21 tables, bibliography, glossary of terms, list of abbreviations, list of author's publications related to the thesis. The used literature includes 151 titles, the list of author's publications includes 10 articles.

Chapter 1 introduces the problem of the integration, analysis and classification of biomedical data, as well as the development of intelligent information systems related to the discussed set of problems. This chapter defines the meaning and relevance of the work, as well as the goal and detailed tasks of the thesis.

Chapter 2 presents a thorough review and comments on the theoretical rationale and analysis of the state of the art for the integration, analysis and classification of biomedical data. The main focuses are on data storage approaches, data integration methods, semantic data integration, software aspects of service-oriented data integration, biomedical data integration and knowledge extraction.

Chapter 3 presents the methodological aspects of the detailed tasks of the thesis, focusing on the formalization and methods for intelligent integration, analysis and classification of biomedical data. The properties of the designed and developed system for semantic integration of data from cancer diseases, aside with forecasting the development of the disease with the help of machine learning methods are considered in detail. Presented are methods for prediction of protein structures and analysis and classification of antimicrobial resistance of metagenomic data. Sequencing data compression methods based on coding noise protection are presented. The integration of data and methodology for creating an advisory system based on ontologies for offering a diet in patients with diabetes is also presented.

Chapter 4 is devoted mainly to the results of the approaches presented in Chapter 3 for solving the tasks, set in the thesis. The main attention in Chapter 4 is paid to the presentation of software solutions for practical application of the developed approaches. It is especially important to emphasize that a holistic approach for developing of a platform for providing software as a service used for the implementation of all systems in the thesis is presented.

Chapter 5 presents the contributions to the thesis, as well as the opportunities for future development.

Objective and tasks of the thesis

The main aim of the thesis is to create a methodology and its practical implementation for intelligent integration of biomedical data and their analysis, using tools of informatics, information technology, bioinformatics and artificial intelligence.

The main tasks related to the purpose of the thesis could be summarized as follows:

1) Tasks related to joint use of data with clinical and laboratory, molecular profile:
   a) Development of a model and software implementation for integration of heterogeneous biomedical data;
   b) Development of a model and software implementation for semantic integration of cancer data;
   c) Development of a model and software implementation for predicting the survivability of cancer patients using machine learning tools
2) Tasks, related to the development of an ontology based decision-making system for diet recommendations of patients with diabetes:
   a) Development of a model and software implementation for knowledge discovery from semantically integrated data for the purposes of an ontology-based decision-making system for use in the treatment of diabetes;
   b) Development of a model and software implementation for generating tips for forming diets based on derived patterns in patients with diabetes.
3) Tasks related to the analysis of metagenomic data for the purpose of classification by microbiome resistance:
   a) Development of a model and software implementation for semantic integration of data from metagenomic research related to pollution in the urban environment;
   b) Development of a model and software implementation for determining the origin, analysis and classification of metagenomic data for microbiome resistance using machine learning methods.
4) Tasks related to alternative methods for predicting the structure of proteins:
   a) Using machine learning methods and software implementation for the purpose of determining the accuracy of folding in protein structures
   b) Assessing the accuracy of the developed methodology for predicting the folding of protein structures using machine learning methods
5) Tasks, related to data generation from parallel sequencing (Next Generation Sequencing, NGS):
   a) Development of a model and software implementation for compression of large arrays of sequence data using noise protection coding algorithms;
   b) Evaluation of the accuracy of operation of the created model for compression of data from parallel sequencing.
6) Development of a platform for providing software as a service used for the implementation of all systems in the thesis.

Structure of dissertation

In Figure 1.2. The structure of the dissertation is presented as semantically related activities through the separate chapters of the work: described and statement of the problem (Chapter 2), methodology of development of the set tasks (Chapter 3) and realization and discussion of the set tasks (Chapter 4). The very semantic part of the development of the idea of semantic integration is presented through the individual activities to the respective chapters, all stemming from the main goal and the corresponding proposal for creating models for data storage and integration based on implemented distributed software solutions as a service. The main idea is to achieve the goal of creating a methodology and its practical implementation for intelligent integration of biomedical data and their analysis, using tools of informatics, information technology, bioinformatics and artificial intelligence.

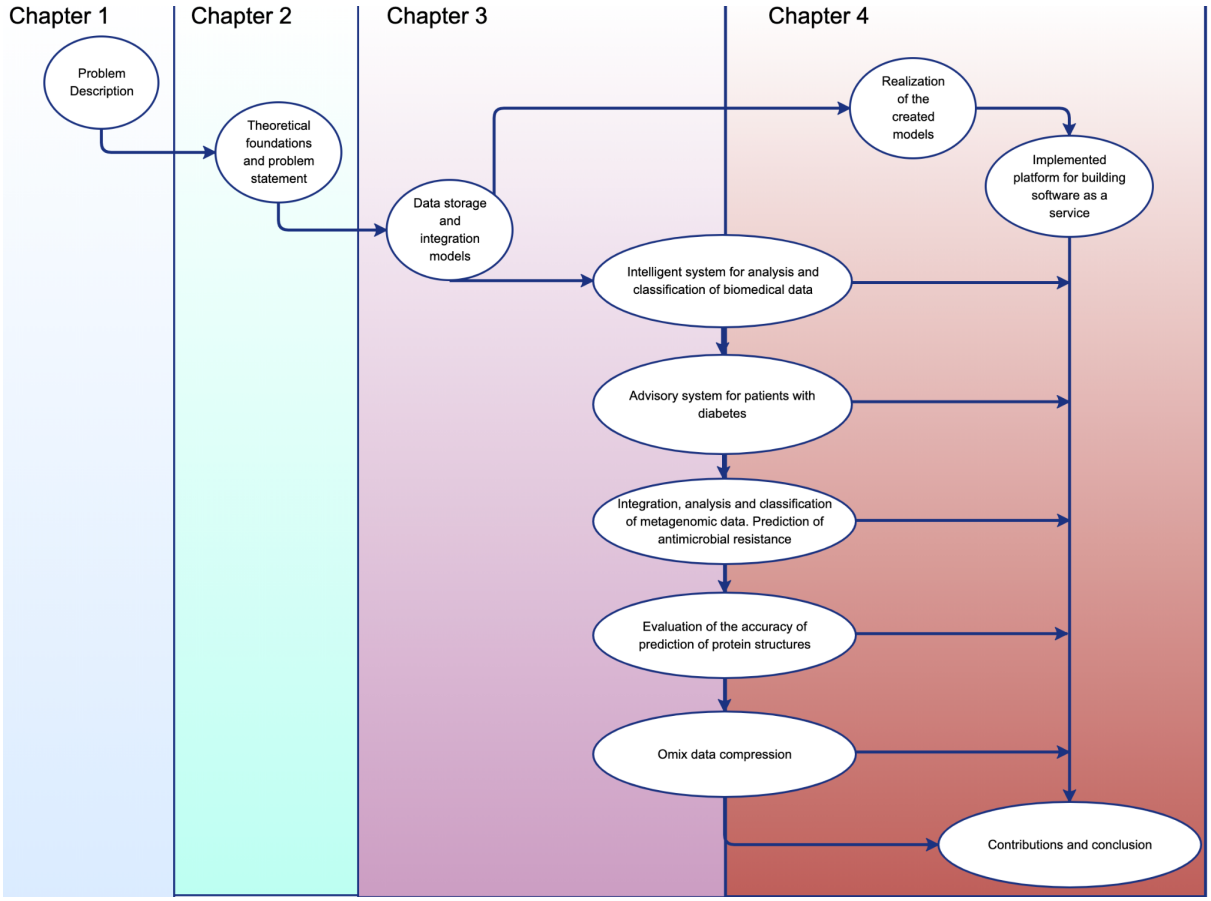


Figure 1.1. Structure of the dissertation

Chapter 1: Introduction

Problem description

With the rapid development of high-performance technologies generating so-called "-omics" data (data from all research and analysis in the field of gen*omics*, prote*omics*, metabol*omics*, transcript*omics*, metagen*omics* and other molecular research domains) in the fields of biology and medicine, especially those of Next Generation Sequencing (NGS) [1], and consequently the rapidly growing volume of such biological datasets, a variety of tools and repositories (databases and web) have been created. servers) to facilitate data management, accessibility and subsequent analysis. A prerequisite for the study of bioinformatics is the ability to search and find, analyze and access data stored in repositories or various resources containing data. For a bioinformatics task, researchers often need to have significant experience with different data sources related to information retrieval, knowledge acquisition and analysis. Undoubtedly, data integration is a time-consuming and resource-intensive process, especially with regard to the import and export of vast amounts of data related to high-performance research and diagnostic technologies in biology and medicine. In this context, the integration of large arrays of distributed, heterogeneous and different in format and time of origin data proves to be a significant problem for the full use of the richness of biomedical data [2]. In this context, the importance of integrating data from biomedical research and practices based on high-performance technologies (such as generating *-omics* data) has two main components [3]:

(1) due to the high level of automation of actual experimental procedures, efforts to obtain the experimental data takes only about 20% or less of the total research effort in a

high-performance omix data generation project; with approximately four fifths of the resources going to the integration and analysis of this data [4];

(2) The answers to the most important, complex biological questions today are rarely provided directly through experimental results, and to provide potential answers, the analysis often involves the integration of diverse data from multiple data sources.

This thesis is devoted to a range of problems accompanying large data sets and their integration in the field of data from medical and biological research and practice. Undoubtedly, the main role in the thesis is played by the ways of data storage, the use of new technologies from NoSQL databases, as well as the subsequent analysis, classification, extraction of knowledge from large arrays of biomedical data. The works for creation of an integrative, ontology-based advisory system for diet in diabetics, analysis and classification of metagenomic data in terms of antimicrobial resistance, and methods for compression of large arrays of sequential data also have a significant contribution to the thesis. A distinctive feature of the thesis is the development of various applications, having new methodological aspects and offering ready-to-use software solutions, which outside the proposed examples are very universal.

The main features for significance and relevance of the thesis could be systematized as:

- Development of a model and software implementation for semantic integration, analysis and classification of biomedical data (both laboratory and clinical) related to cancer;
- Development of a model and software implementation for predicting survival in cancer patients using machine learning methods;
- Development of a model and software implementation of an ontology-based decision-making system for diet recommendation for patients with diabetes;
- Development of a model based on machine learning and software implementation for prediction of protein structures, their classification and analysis to assess accuracy;
- A model and software implementation for data compression from parallel sequencing by means of noise protection coding algorithms have been developed;
- All software realizations in the thesis are made on the basis of a created platform for presenting software as a service.

Complexity of the problem

Themain characteristics of the complexity of the problems before the thesis could be: The

- heterogeneous structure and origin of the data;
- The problems in front of a single theoretical basis, taking into account the specifics of the integration of biomedical data;
- Relatively insufficient classification of biomedical data;
- Non intensive development and use of subject ontologies in medicine and biology;
- Approaches to knowledge acquisition and service creation on this basis are poorly integrative;
- Research related to the use of large data sets is related to new approaches in data storage, transfer, compression.

Research domains and means

Informatics and Computer Science. The main task of the work - data integration is one of the most in need of fast and universal solutions to problems in the field of informatics and computer science. The use of NoSQL databases for storage and integration of data used in the work is also a rapidly evolving field of informatics. The use of machine learning methods in almost all tasks in the thesis undoubtedly contributes to obtaining optimal solutions. The software realizations of the individual tasks in the thesis are developed in a modern environment of programming languages, libraries, platforms and user interface.

Bioinformatics. Bioinformatics is an interdisciplinary scientific field that deals with the study of data from different parts of biology and medicine with the help of informatics.

Bioinformatics is determined entirely by the development of algorithms and software that can be used for integration and analysis of incoming information, as well as for performing numerical experiments, classification and knowledge extraction. With the development of information technology, as well as with the development of technology in the field of medical and biological research, the volume of data generated is growing very fast, sometimes faster than the computing resources themselves, which greatly increases the possibilities for expanding research on achieving new much more significant results, as well as extracting new knowledge.

Artificial Intelligence. Artificial intelligence is an area that is used to build intelligent software systems that can solve a growing range of problems that have so far been done with conventional computing tools. This range of problems can include any tasks that cannot be solved with formally described algorithms - data analysis, image and speech recognition, robotics, self-learning programs. Artificial intelligence is an area that is developing extremely fast and with successful applications in all areas. It has a special role in the work with biomedical data and in general, as an application in medicine and modern biological research.

Chapter 2: Theoretical foundations and analysis of the situation on the problems of integration, analysis and classification of biomedical data.

Data integration systems are characterized by an architecture based on a global scheme and a set of sources. The sources contain real data, while the global scheme provides a coherent, integrated and virtual view of the main sources. Therefore, modeling the relationship between sources and the global scheme is a key aspect. To this end, there are several basic approaches for data integration that can be roughly classified into five groups [23], [24]: data storage, database aggregation, service-oriented integration, semantic integration, and wiki-based integration. . In all these groups, an increasingly important component of data integration is occupied by activities for the development of various ontologies, to more specifically address the technical challenges of defining descriptors and identifiers of information to be shared and integrated by different resources [19], [25].

Nowadays, huge amounts of data are collected from many diverse sources, which generate real-time data with different qualities - which is called *big data* .

We can say that the integration of large data sets differs from the traditional integration of data in many dimensions: volume, speed, diversity and reliability, which are the main characteristics of large data: Data volume is the original attribute of the so-called large data sets. data. Today, the number of connected devices and people is higher than before, which greatly affects the number of data sources and the amount of data worldwide and in each area. The variety of data sources suggests that we have more diversity in the formats in which the data is stored. We have structured and unstructured data at a high level. In each type we have a huge number of formats: text, images, sounds, documents, spatial data and others. In fact, we have different data quality, which leads to a problem with their reliability. We can find uncertain or inaccurate data in all areas from social networks to bioinformatics systems that integrate data.

Data storage

Data storage approaches offer a solution on how to do this (conditionally in one place) in order to facilitate access to and management of a wide variety of data from different sources. Data warehouses focus on translating data, collecting all available data from many different sources, transforming and adding it to the repository itself. The data warehouse has many aspects and can be placed in different physical locations, depending on the volume, speed of generation and variety of data to be integrated. A data warehouse is a collection of many different databases that can interact with each other. The main properties of data warehouses range from multidimensional analysis to statistical and data retrieval requirements to exploration capabilities, as well as the introduction of adaptive analytical applications, these

technologies being part of a stable and proven environment.

The main properties of the data warehouse are: not to contain redundant data (nonredundant), the storage to be always available (stable), the data stored in the storage to always be exactly the ones that the user has chosen (consistent). All these properties can be satisfied by cloud services, where the resources available are much larger and the possibilities for building a hierarchical data warehouse are possible.

The challenges of storage are mainly related to the volume, speed and variety of large data sets. Storing large data sets such as traditional physical storage is problematic because hard disks (HDDs) are often damaged and traditional data protection mechanisms (eg RAID or an array of independent disks) are not effective when storing on a petabyte scale [26]. It is necessary to develop principles and algorithmic solutions, taking into account the spatio-temporal models of data use, to determine the analytical value of the data, as well as the relevant data for their preservation by balancing the costs of data storage and transmission with the rapid accumulation of big data [28].

Data transfer

Data transfer takes place at different stages of the data life cycle, as follows: (i) data collection from sensors, biological instruments; (ii) integration of data from multiple data centers; (iii) data management for the transfer of integrated data to processing platforms (eg cloud platforms) and (iv) data analysis to move data from storage to an analysis host (eg High Performance Computing clusters). The transfer of large amounts of data poses obvious challenges at each of these stages. Therefore, intelligent preprocessing techniques and data compression algorithms are needed to effectively reduce data size before data transfer [29].

Resilient data management

It is difficult for computers to effectively manage, analyze, and visualize large, unstructured, and heterogeneous data. The diversity and reliability of big data redefine the data management paradigm, requiring new technologies (eg Hadoop, NoSQL) to clean, store and organize unstructured data [30]. While metadata is essential to the integrity of the origin of the data [31], the challenge remains to automatically generate metadata to describe the big data and related processes. Generating metadata for geospatial or biological data is even challenging due to the inherent characteristics of high-dimensional and complex data (e.g., space-time correlation and dependence). In addition to generating metadata, large data sets also pose challenges to database management systems (DBMS), as traditional RDBMSs lack scalability for managing and storing unstructured big data [32], [33]. While non-relational (NoSQL) databases such as MongoDB and HBase are designed for large data sets [34], models of flexible adaptation of NoSQL databases for processing geospatial or biological big data by developing efficient spatio-temporal indexing and query algorithms are increasingly they are still a challenge [35]. NoSQL solutions are based on three main theorems: CAP, BASE, and a possible sequence.

Data analysis

Data analysis is an important phase in models and algorithms that use large data sets to extract information and patterns [45]. Big data analysis, in turn, poses the following problems for the complexity and scalability of basic algorithms [46]. Big data analysis requires complex scalable and interoperable algorithms [47] and is addressed through analysis programs and parallel processing platforms (eg Hadoop) to use the power of distributed processing. However, this divide-and-conquer strategy does not work with the multi-layered and multi-scale iterations [27] that are required for most biological data analysis algorithms. In addition, most existing analytical algorithms require structured homogeneous data and have difficulty in processing, taking into account the heterogeneity of the big data [29]. This open question requires either new algorithms that deal with heterogeneous data or new data preprocessing

tools to make them structured to fit existing algorithms. In bioinformatics, optimizing existing algorithms for spatial analysis by integrating space-time principles to accelerate the discovery of biological knowledge is a challenge and has become a high-priority research field of "bioinformatics thinking, calculations and applications" [27].

Data architecture

Big data is gradually transforming the way research is conducted, as evidenced by the increasingly data-driven and open scientific approach [47]. Such transformations pose challenges to the system architecture. For example, the seamless integration of different tools, geospatial services and biological research remains a top priority. Additional priority issues include integrating these tools into reusable workflows, incorporating data into functionality promotion tools [45], and sharing data and analysis between communities. The ideal architecture will seamlessly synthesize and share data, computing resources, network, tools, models, and most importantly, people. Bioinformatics cyber-infrastructure is actively used in biomedical sciences [51]. NCBI, although still under development, is a good example of such a cyber infrastructure in the bioinformatics field.

Service-oriented integration

Data storage and aggregation focus on centralizing data access through data translation and query translation, respectively. They face some problems arising from data storage and processing, frequent updates and high costs for data exchange and / or maintenance. In part to avoid these problems, there is an improved and decentralized approach in which individual data sources agree to open their data through web services (WS). WS are designed to communicate between computers over a network and are described by the Web Services Description Language (WSDL). There are several different protocols for WS, such as SOAP (Simple Object Access Protocol; protocol for exchanging XML-based messages over computer networks), REST (REpresentational State Transfer; a simple protocol implemented using HTTP methods). The WS supports computer-to-computer interaction through a Web application programming interface (Web API) and can execute a database request or calculations. In the context of data integration, data can be accessed programmatically via WS and data sources serve as service providers. Therefore, such an approach can be considered as a service-oriented approach.
The service-oriented approach includes data integration through communication between computers via a web API and up-to-date data retrieval from various data sources. This approach remains a challenge, mainly because its success in heterogeneous data integration requires many data sources to become service providers by opening their data through WS and by standardizing data and nomenclature identities to facilitate exchange and analysis. of the data.

Semantic integration

Most web pages in biomedical data sources are intended for human reading (eg HTML). The Semantic Network [65], [66] aims to describe the data in a way that computers can understand and to build an interconnected network through which computers can easily and unambiguously interact and analyze. According to the World Wide Web Consortium (W3C) definition statement, the goal of the Semantic Network is to create a universal medium for data exchange using several standards, including the Resource Description Framework (RDF; http://www.w3.org/). RDF), RDF schema (RDFS - language for describing the RDF dictionary; http://www.w3.org/TR/rdfschema), web ontology language (OWL; http://www.w3.org/owl) and standard language for web queries SPARQL (http://www.w3.org/TR/rdf-sparql-query) for RDF. RDF provides standard formats (eg XML format) for data exchange and describes the data as a simple statement containing a set of triplets: subject, predicate and object. Any two statements can be related by the same subject or object. OWL is based on RDF and Uniform Resource Identifier (URI) and describes the structure and meaning of data based on ontology,

which allows for automated thinking of data and conclusions from computers. The semantic network provides a machine-readable way of presenting data and interoperability [67], [68].
The application of semantic web technologies for the integration of biological data is a significant advance for bioinformatics, allowing automated data and knowledge processing. Semantic integration uses ontologies to describe data and thus represents ontology-based integration [68]. The semantic network continues to evolve and its application in the integration of biological data has several limitations. Semantic integration locally stores a large collection of RDF documents by copying data from multiple data sources and converting data to RDF format.

biomedical data standards

High-performance technology platforms, such as parallel genomic sequencing (NGS) platforms in biomedicine, can generate vast amounts of data in a relatively short period of time. To keep up with the revolution in sequencing technology, genome sequencing projects are gradually shifting from classical model organisms (eg, fruit fly (*Drosofila M*), mouse, yeast). On the other hand, it is impossible to integrate such large amounts of data into one environment (eg data warehouse). Data sources are designed for different purposes and perform different functions. Therefore, it is promising to create an efficient way to exchange data between these distributed and disparate data sources. However, a dozen data sources are intended only for data storage, but not for data exchange. The growing volume of biomedical data also requires "computer readable" approaches to data integration. To facilitate data integration, data sources need to become service providers. In other words, data sources must not only serve as data providers that provide human-readable data with web interfaces (eg HTML), but also function as service providers that provide computer interoperability data through WS. Service providers provide data such as WS, facilitating the interaction between computers and thus allowing automated integration of data from multiple data sources [74]. As mentioned, there are several different protocols that can be used to create a WS. Among them, SOAP and REST are widely accepted (Figure 2.9.). SOAP is a well-defined standard with XML-structured request and response messages, while REST is relatively light, relying on HTTP methods (namely POST, GET, PUT or DELETE). Most commercial applications display their services as the RESTful Web API (Figure 1), largely due to its simplicity and ease of implementation.

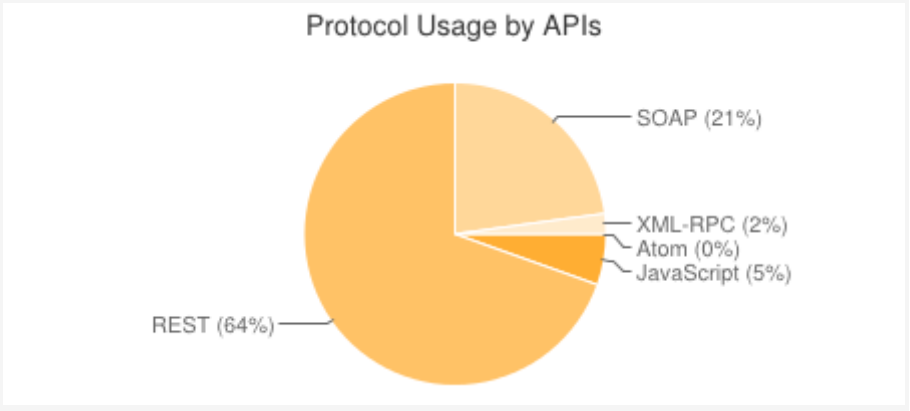


Figure 2.1. Statistics of various Web API protocols

It is equally important that data integration also requires standardization of nomenclature and ontologies for biomedical data.

Software implementation for data integration

The purpose of data integration is to enable the automatic combination of information from different resources without human intervention, so as to cope with the growing

accumulation of biomedical data. To this end, the data to be integrated must be redefined more broadly, which includes not only sequences and other raw data, but also methods, tools, algorithms, analyzed results, open knowledge (see Article on knowledge integration). ;, 2007) and even relationships between people [77]. All types of data can be provided as a service. That is, the raw data must be accessible through WS, the methods, tools and algorithms used for data analysis must be offered as WS (ie SaaS, Software as a service), and the analyzed results and the discovered knowledge must also be available. to be supplied as WS [77]. As a result, WS performs various data manipulations, including data retrieval, integration, analysis, visualization, and sharing.

A pipeline with a combination of multiple WS can achieve data integration. Such WS-based pipelines reduce technological barriers and provide users with a lighter programming environment. WS-based pipelines, which include computer-to-computer data exchange, simplify data integration and analysis, maximize sharing and reuse, and function as a connecting environment for users with similar research interests, and finally to form scientific social and project community.

Integration of biomedical data and knowledge discovery

The integration of multi-omics data from the same set of objects is expected to increase the accuracy and speed of predicting results, such as early detection of cancer based on data from many platforms. Omics data analysis technologies feature high-performance interfaces that facilitate the study of the genome, epigenome, transcriptome, proteome, and metaboloma in a globally unbiased manner. Approaches to analyze -omics data are now used to understand complex biological systems and to reveal the molecular signatures that underlie complex phenotypes [80]. When integrating multi-OMICS data that is not indexed, other approaches are used, which are part of the machine learning models in order to discover new groups and subgroups of objects. Determining the type of cancer or tumor typing is a common problem that is solved through machine self-learning without a teacher. Clustering efficiency can be quantified using simulation studies, extensions of several views of index criteria, and enrichment analysis [81].

Integrating cancer data, which are taken from different clinical laboratories, have different formats, properties and structure, but describe the same cancer. In this case, the methods of machine learning are used to extract regularities and similar properties between the individual objects. In this way, links between heterogeneous data can be established.

In the analysis of large biological data, it becomes mandatory to examine the basic principles of data integration from multiple sources in order to provide a higher level view in order to derive new knowledge based on the use of machine learning methods in data integration. [82] Over the last decade, high-performance technologies have been widely used in conjunction with clinical trials to investigate various diseases to decipher basic biological mechanisms and to develop new therapeutic strategies. The generated high-performance data often correspond to measurements of different biological parameters (eg gene expression, RNA transcripts, proteins), represent different views of the same object (eg genetic, epigenetic) and are created by different technologies (eg microarrays), next generation sequencing, etc.). The data are diverse, of different types and formats.

Chapter 3: Formalization and methods for intelligent integration, analysis and classification of biomedical data.

Heterogeneity of biomedical data. Over the last decade, high-performance technologies have been widely used in conjunction with clinical trials to study various diseases to decipher the basic biological mechanisms and to develop new therapeutic strategies. The generated high-performance data often correspond to measurements of different biological units (eg transcripts, proteins), represent different views of the same object (eg genetic, epigenetic) and

are created by different technologies (eg microarrays, RNA sequencing) . The data are diverse, of different types and formats. There is an obvious need to integrate data in order to store, access, link, analyze and retrieve easily [86].

In summary, from a methodological point of view, the integration of data in the context of bioinformatics means the integration of data from different sources and the purpose of creating a unified idea, form, mode of entry and the possibility of deriving knowledge. Each data source in bioinformatics has its own approach to data structuring, which increases the complexity of integration. Data integration and biomedical data analysis are separate disciplines that have developed relatively in isolation. There is a general consensus that combining the two disciplines in order to develop more sustainable methods of analysis is necessary [87], [88]. Data integration mainly involves queries to various data sources. These data sources may be, but are not limited to, separate databases or semi-structured data sources distributed over a network. Data integration facilitates the division of the entire data space into two main dimensions, relating to where the data or metadata knowledge is located and to the presentation of data and data models. Biomedical experiments use a large number of different analytical methods that facilitate the extraction of relevant knowledge from distributed information.

Methods for integrating biomedical data. The heterogeneity of biomedical data makes any integrative analysis highly challenging. Data generated with different technologies includes different sets of attributes. When the data are highly heterogeneous and poorly connected, two interrelated integrative approaches are applied: horizontal and vertical integration (Figure 3.1.). Horizontal data integration brings together information of the same type, but from different data sources and, potentially, in different formats. This facilitates the aggregation of disparate data, such as clinical information, from many different sources into a single data model. Vertical data integration, on the other hand, means linking different data types to achieve better analysis and derivation of knowledge for multiple data types. This approach helps to manage the links between the patient's genetic expression, clinical information, available chemical knowledge and existing ontologies. Most existing approaches to data integration focus on one data type or one disease and cannot facilitate the integration of a cross-type or disease [91], [92].

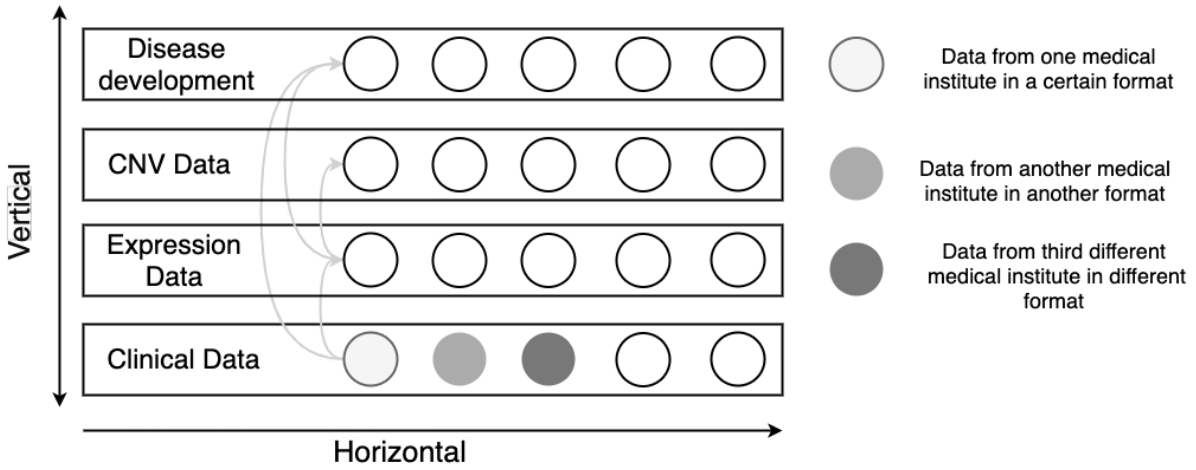


Figure 3.1. Vertical and horizontal integration of data from different sources and with different formats.

The main problems facing vertical and horizontal data integration are related to data heterogeneity. Data heterogeneity is a general concept that in the context of data integration can be defined as a set of problems. One of the main problems that lead to heterogeneity is the use of different operating systems, platforms and hardware configurations. In the context of biomedical data, there is also a very strong heterogeneity in the used data presentation formats. Different formats of the same data lead to heterogeneity in the structure and conclusions that can be drawn from these data. The differences in knowledge that can be deduced are called semantic heterogeneity. It is characterized by the use of many resources, which by analyzing

from different entry points are extracted contradictory knowledge. One common solution to this problem is to use ontologies.

Semantic integration of biomedical data.

The basic idea of the semantic network is to add machine-readable metadata to resources in the global network to define and describe the relationship between them. Semantic web technologies are able to assimilate this acquired information. In addition, they do not build a separate network, but function as an extension of the current network. Semantic Web technology consists of the hierarchical use of different standards and technologies, in which each layer uses the capabilities of the layers below.

The problem with using a purely semantic integration approach is the introduction of multiple metadata. These metadata must also have a common format, which is a big problem when working with biological data. These metadata should be able to describe sufficiently well, for example in clinical terms, cancers and characteristics of bacteria and other related diseases. For this purpose, a dynamic data linking model is needed, which can be adapted to the type of data being integrated. For this purpose we can use the so-called Linked data. Linked data is the most advanced form of publishing data on the web according to W3C specifications. The W3C describes a number of good practices for publishing data on the web [94]. This approach only makes sense when the data is publicly available. This data network is sometimes called "Data Network", a term with a more practical emphasis than the older but equivalent "Semantic Network".

The basic idea of related data is that any object - animated or inanimate, private or abstract - can have an identifier: a universal resource identifier or URI (Uniform Resource Identifier). Data in the data network refers to objects that are very accurately identified. URIs are sequences of characters with several parts separated by periods and slashes. For example, Uniform Resource Locator URLs, which are web addresses entered into a web browser to produce a page, are also a type of URI. This match, which turns URIs into a superset of URLs, is not accidental: the expected behavior when entering a URI in a web browser is to retrieve information about the identified object. The string of symbols used to identify something retrieves more information about that thing. URIs aim to name all objects in the world in the same way.

The second key idea of related data is that information can be given for each identifier. The data for the selected object is retrieved from the specified URI. Based on it, a division of the used data sources can be made. In this way, the distribution of data and the organization of data that are common are achieved.

Integrate data from many different sources on the example of cancer. In the thesis an approach is proposed, for horizontal and vertical integration of data whole sets of data are united, as the semantic integrity of the data is preserved and enriched. By combining data from multiple cancers, a data network is thus created where objects, such as proteins, clinical features, and expression characteristics, are interconnected [95]. Data can often be represented as networks, where nodes show biologically significant objects (usually genes or proteins), and arcs represent connections between these objects (eg regulation, interaction). In the generated network, nodes represent patients and arcs represent similarities between patient profiles, consisting of clinical data, expression profiles, and copy number variation (CNV) information. Such a network can be used to group similar patients and to associate these groups with different characteristics [96]. The main challenges here are: (1) building an appropriate interconnected data network, finding the semi-structure of the data model [97] and mapping the claims of the applied data integration model [98]; and (2) data cleanup combined into a formal data integration workflow Figure 3.3.

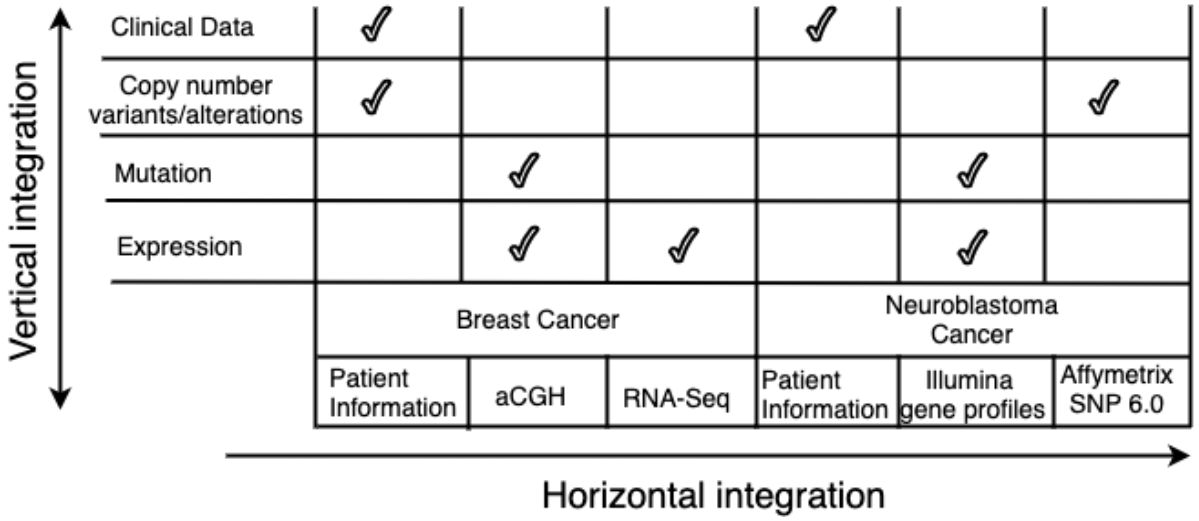


Figure 3.3. Integration of heterogeneous data. The horizontal axis shows different types of data and sources integrated for a specific patient. Along the vertical axis, integrated data types related to the studied cancers and related to a specific patient are given.

Horizontal and vertical data integration requires different databases, as each of these approaches addresses different aspects of the integration problem. Horizontal data integration deals with unstructured and heterogeneous data. Thus, a document-based database (such as MongoDB) is used in the work, which can process different types and formats of data. For vertical data integration, a database based on graphs is used, as it is suitable for presenting relations that are crucial in this case. In this study, all relationships are established between existing records for each object and are represented by a semi-structure.

The proposed integrative framework facilitates direct data analysis. We first focus on a specific clinically relevant application: modeling and predicting the survival time of cancer patients. This consists in the application of both conventional classification methods and machine learning algorithms. Through data integration, a new integrated and universal, ie applicable to both cancers, survival time prediction function is introduced. This feature is made up of three clinical features that are most related to survival. In addition, this integrated feature provides a connection to the newly developed connected data network. This feature is used in the method of conventional classification of k-neighbors to find patients who are most closely associated with research. Then, through the associated data, we find other patients who may not have the new integrative function, but are still associated with different types of data, such as gene expression or CNV. Machine learning models based on regression vector support and decision tree, survival time prediction and cross-validation are then used.

Developed methodology for the semantic integration of biomedical data from different diseases.

The task is to provide a method based on related data and open source technologies that combines knowledge from many existing open sources to effectively integrate raw libration data. Raw libration data that can ultimately be integrated to fully elucidate complex phenotypes include functional gene annotations, gene expression profiles, proteomic profiles, DNA polymorphisms, DNA copy number variations, epigenetic modifications, and more. [101].

A specific challenge in the study is to integrate and analyze sets of unbalanced and unstructured data. Molecular data is in raw format with all fields and attributes generated by sequencing technology or microchips. Before starting the integration process, it is necessary to perform some operations for pre-processing of raw formats and to generate an appropriate new

data structure. In this case, we are working with data sets rich in links, and it is essential to be able to find many annotations about existing links that will help improve the set of links through appropriate resources from the available sources of knowledge.
The thesis paper has chosen an approach based on semantic integration, as most of the characteristics of the data used have different semantics for each patient, which is an essential background for personalized medicine. Expected results of this type of approach include the identification of latent protein subtypes, characterized by common network change patterns and a predictive cancer development model based on knowledge of the fused proteins.

Description of the problem area and data. Data include functional gene annotations, gene expression profiles, proteomic profiles, DNA polymorphisms, variations in the number of DNA copies, epigenetic modifications, and more. [19]. The raw data in each studied data set are in a certain format and have specific semantics. The field (attribute) in each data set has different meanings due to the technology and the subsequent record. The data provided by themselves also contain information on mutated proteins, expression and CNV.

The initial point for transformation, grouping and integration are the patient files. The generated record for each specific patient contains attributes such as age, gender, nationality, etc. Two data sets were used in this study - neuroblastoma (NB) and breast cancer (BC). The neuroblastoma kit contains RNA-Seq gene expression profiles of 498 patients, as well as Agilent microchip expression and aCGH data for a concomitant subset of 145 patients and relevant clinical information. The breast cancer data set contains microchip profiles and copy number data and clinical information (survival time, multiple prognostic markers, therapy data) for approximately 2,000 patients.

Main characteristics and innovations in building the methodology for semantic integration.

The data integration approach that was developed was initially application-specific - to combine data from real studies and treatments of neuroblastoma and breast cancer - but its design features make it sufficiently generic and applicable in a wide range of thematic areas. As a result of its application, different data sets are combined, and the semantic integrity of the data is preserved and enriched. In a specific case, by combining data from multiple sources (Fig. 3.4), a new data network is created where objects, such as proteins, clinical characteristics and expressive characteristics, are interconnected [90]. In this network, nodes represent patients and arcs represent similarities between patient profiles, consisting of clinical data, expression profiles, and CNV data. Such a network can be used to group patients and to associate these groups with different characteristics. The main challenges here are: (1) building an appropriate interconnected data network, finding the semi-structure of the data model and mapping the claims of the applied data integration model [106]; and (2) data cleanup combined into a formal data integration workflow.

Using different databases for horizontal and vertical data integration. These different databases are needed because horizontal and vertical data integration address different aspects of the integration problem. Data on horizontal data integration are unstructured and heterogeneous. In this way, a document-oriented database is used, which can process different types of data and formats. For vertical data integration, database graphics are used, as it is suitable for representing relationships - crucial in this case. In this study, all relationships are established between existing records for each object and are represented by a semi-structure.
An integration model based on a NoSQL database can potentially combine medical research data, as an alternative to the most commonly used methods for statistical analysis and machine learning. Most NoSQL database systems share common features that support scalability, availability, flexibility, and provide fast access time for data storage, retrieval, and analysis [107], [108]. In addition, the potential of the model can be extended by using multiple data sets, regardless of the level of heterogeneity, specific formats, data types, etc.

The methodology for integrating unstructured data from test samples and patients is based on the proper use of non-relational databases and domain ontologies such as gene

ontology (GO) [112]. The data included in the study contained hidden links between proteins provided by different patients in studies of the two diseases (BC and NB). All available information on already established relationships in data sources is used, searching for and finding additional information in some third party sources, to achieve semantic integration of data. In this way, step by step, a network is developed that combines the protein connections between patients and diseases. The challenge here is to preserve all relationships with their cyclical dependencies. The latter are possible because one patient has a link to mutated protein (s), mutated protein (s) have a reference (s) to expression, and other patients have references to the same protein (s).

For each disease (BC and NB), each patient has a different set of mutated or expressed proteins. Only small sets of mutated proteins are equal and exist in each patient. All proteins belong to families that contain many related proteins. By applying semantic annotations and search techniques, all proteins that are semantically related to the studied diseases are detected and combined. In this way, all available and necessary information for a similar number of related proteins can be found.

During the analysis of raw data, a kind of "semi-structure" of the data is created - a structure containing only the attributes existing in each record. In the so-called semi-structured data objects belonging to the same class (protein mutations expressing transcripts and CNV) may have different attributes, although they are grouped together, and the order of the attributes is not important. Semi-structured data are becoming more common, e.g. in structured documents and when performing simple data integration from multiple sources. Traditional data models and query languages are inappropriate, as semi-structured data are often irregular: some data are missing, similar concepts are presented using different types, heterogeneous sets are present, or the object structure is not fully known [103].

Models for extracting knowledge from biomedical data. The data used in the thesis are based on clinical records, including: age of the patient, stage of tumor development, tumor size and life status of the respective patient. There are also data on different types of therapies and surgical interventions, thus including many characteristics related to the development of the disease - tumor size, age at diagnosis, tumor stage, information on surgery and treatments such as chemotherapy, hormone therapy and etc. in the raw data. Two sets of data from breast cancer studies are used in the present work. The first data set contains profiles of 498 patients as well as relevant clinical information. The second set of breast cancer data also contains genomic profile data and clinical information for 2,000 patients. The different types of available data and sources of information are shown in Figure 3.6. The same type of information is provided from different sources in different formats. We integrate all data both horizontally and vertically.

Primary data processing. The database consists of two layers: first, a non-relational document-oriented database - a class of databases that store their data in the form of documents. These databases are horizontally scalable and much more flexible than relational databases. In addition, the second layer is a graphical database - a class of databases that store data in the form of a graph and use a technique called meaning without an index. In the graphical database, the main emphasis is on the relationship between the data. In a document-oriented database, a constraint (called a "data schema") is applied based on the generated semi-structure. The attached data schema for each record for each data type combines data in different formats and from different sources. For each data type, this data scheme always contains an ID and a sample ID (representing the name of the subject as provided in the clinical information).

The semi-structure approach is used to integrate all disparate data. In this way, horizontal data integration is performed. For vertical data integration, two layers of the semi-structure are used - for each data type (containing only attributes that exist in each record) and for all data types (containing ID and Sample ID). This creates a network of connections between all types of data to manage them.

Data integration. As noted above, by definition, data integration is the process of combining data of different types and from different sources and combining them into meaningful and

valuable information. For data integration we use the newly created network of relationships. In these networks, the nodes represent patients and the edges represent similarities between patient profiles. The similarity means that two patients are linked by multiple proteins. These networks of connections can be applied to groups of patients and connect these groups with different clinical characteristics. The network has two layers. The first layer, covering internal relationships, is made up of relationships generated by raw data. The raw data contained a description of each patient, with associated expression data, copy number variants, and clinical information. This information is transformed into relationships between patients and expressed proteins. The second layer is based on semantically related data from external sources of knowledge. These sources provide information on additional proteins related to those existing in the dataset. These new relationships are stored in the created graphic database. In order to use additional information from external sources of knowledge, they are connected to the network via hyperlinks (URLs). In this way, it can avoid visual incomprehensibility, which would be caused by an excess of information. These two layers are combined in one network with different weights for each connection. The approach to data integration presented in the thesis consists of several steps (Figure 3.7.).

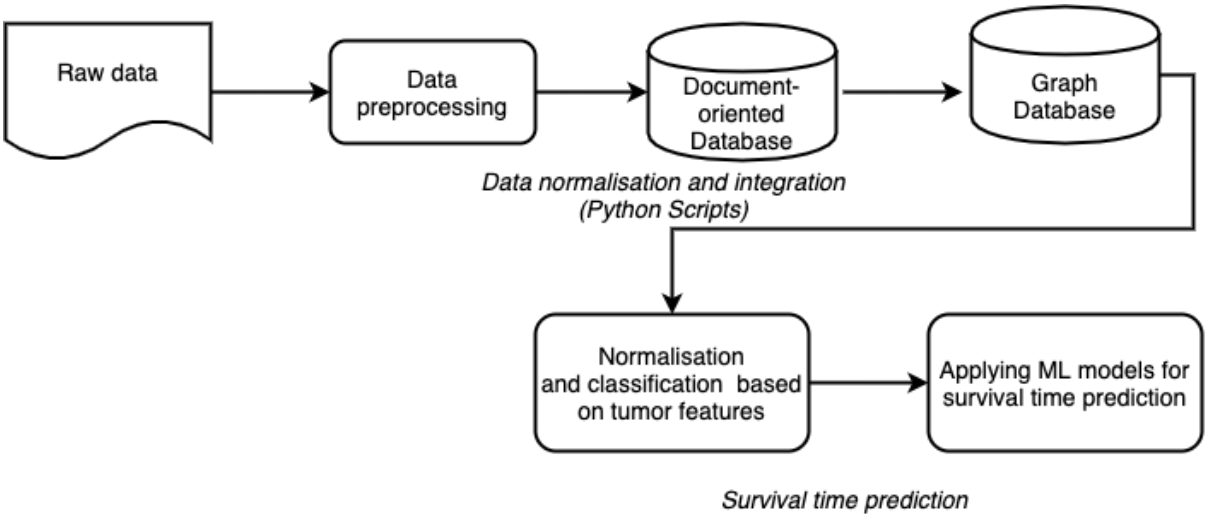


Figure 3.7. Architecture of the proposed approach for data integration

Survival prediction model for cancer patients

Intelligent biomedical data integration systems have two main components, reflected in the thesis: semantic data integration and models for extracting knowledge from these integrated data. In this regard, a model was developed that uses semantically integrated data and predicts the development of diseases, such as breast cancer.

For the purposes of the present study, a new and universal prognostic parameter was developed - tumor-integrated clinical feature (TICF). This characteristic is constructed by numerically combining the tumor stage, tumor size, and age at diagnosis (Figure 3.8) in exactly this order. The order of aggregation of these clinical parameters is important due to the classification of clinical information about tumor development and its importance for the patient's survival rate. In particular, a patient with a stage four tumor will have a shorter survival time than a patient with a stage two tumor. The next characteristic - the size of the tumor - is added secondly, because as the size of the tumor increases, the patient's survival rate decreases. Tumor size is the second characteristic also because it is less important for survival time than the tumor stage. The third characteristic used is the age at the time of diagnosis, at which older patients have a lower survival rate. If the order of aggregation of these components that make up the TICF differs, patients with remote survival characteristics will be incorrectly grouped. This provides a normalized distance between patients, which is essential in our next ML approaches to predict survival time. The applied normalization approach is based on the standard deviation of the

training set so that it can later apply the same transformation on the test set. Alternative standardization is based on scaling the characteristics aimed at obtaining them between a given minimum and maximum value, often between zero and one, or so that the maximum absolute value of each characteristic is scaled to the size of the unit. The motivation to use this scaling includes resistance to very small standard deviations of the characteristics and preservation of zero records in scarce data.

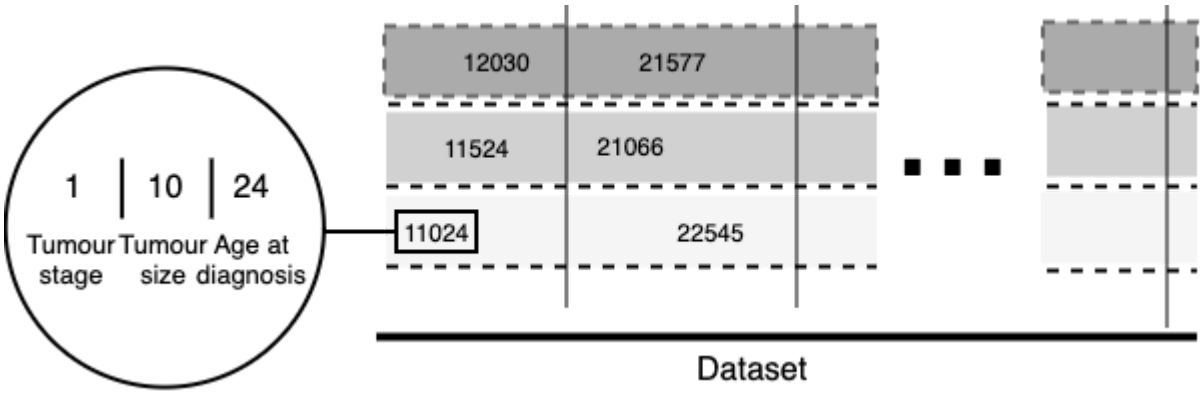


Figure 3.8. TICF forecast parameter

The next step in this methodology is the use of ML models to predict survival times and evaluate them. The ML models used support support vector-based regression (SVR) with different cores: RBF, linear and polynomial, as well as Lasso regression, Kernel Ridge regression, K-quarter regression, decision tree and multilayer reception (MLP) regression. More formally, a machine-based reference vector method constructs a hyperplane or set of hyperplanes in a space of high or infinite dimension that can be used for classification, regression, or other tasks such as finding extraordinary points. Intuitively, good separation is achieved by the hyperplane, which has the longest distance to the nearest point of the training data from each class (the so-called functional margin), which brings a lower error in summarizing the classifier. We also use the Stochastic Descent Gradient (SGD) model, an algorithm for training a wide range of models in ML, including (linear) SVM, logistic regression, and graphical models. When combined with the inverse replication algorithm, this is the de facto standard ANN training algorithm.

Using machine learning to evaluate the accuracy of predicting of protein structures

In this work, ML models are applied to estimate the accuracy of both properties, KB energy and probability, which can be used as scoring functions. Given that probability is a more accurate measure of adjusting the sequence structure. The purpose of this study is to confirm this using ML models. The idea is to check whether the model predictions for the probability values will be more accurate than those of the KB energy. This will show that the probability provides an opportunity for better use of structural information in forecasting the energy of KB.
Cross-validation divides training data into several unrelated cohorts of approximately the same size. Each cohort in turn is used as test data, while the other cohorts are used as training data. The prediction model built on the training data is then applied to predict the labels of the test data classes. This process is repeated until all cohorts are used as test data once, and then the accuracy of the predictions of all blind tests are combined to obtain an overall performance assessment.
Methodology for energy accuracy assessment. For the purposes of this study, an ML-driven approach was developed to assess energy accuracy (E) and probability based frequency (L) to predict the structure of a knowledge-based protein. Both approaches are based on statistics on the latent / exposed properties of residues. The sequence and pattern of each of the 245 protein objects are transformed into numerical values that can be used as parameters in ML models.

To predict the energy values and the probability of KB, three controlled ML prediction models were used. The selected models are from the python scikit-learn package: 1) Lasso - linear_model (alpha = 0,1), 2) Regression of the nearest neighbors (NNR) - kNeighborsRegressor (n_neighbors = 5, algorithm = 'kd_tree'), 3) Decision tree regression (DTR) - DecisionTreeRegressor (max_depth = k). For each of the models, a k-fold cross-check is used to divide the set into k smaller sets for better evaluation.

Linking data for ontology based decision support system for patients with *Diabetes Mellitus* type 2.

In the presented study, the principles of linked data are applied in the proposed methodology aimed at using all existing information, together with available ontologies developed specifically for diseases of interest, such as DMTO, which is a newly developed OWL 2 ontology containing complications related to with diabetes, symptoms, medications, laboratory tests. The new DMTO updates address the options for customizing treatment plans for patients with type 2 diabetes. The patient data used in the study are from clinical laboratory test records and the result presented may serve as an integral part of an interoperable information system. for medical documentation.

Created methodology. The DMTO ontology is used to manage and store patient data, as well as to support the creation and delivery of diet plans based on the patient's laboratory tests. The subject knowledge base is applied using DMTO and SWRL rules. Specific rules are added to calculate the amount and proportions of macronutrients that the patient should take based on the results of laboratory tests. These rules are used to make dietary treatment plans through official information and knowledge. After importing a set of patient data, an appropriate formal argument is made and a treatment plan is generated (Fig. 3.9.).

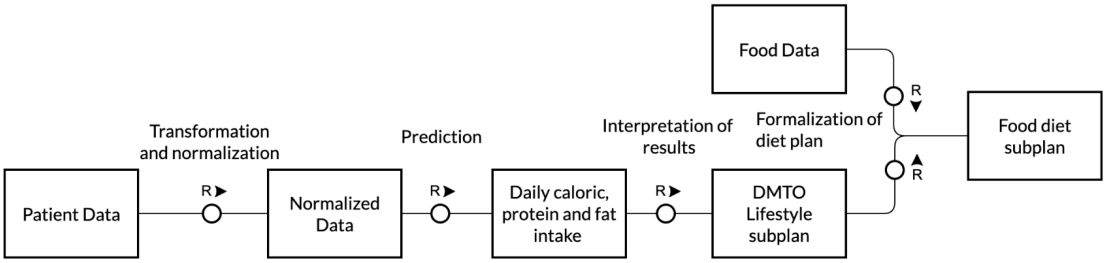


Figure 3.9. The process of generating a diet plan.

The first and most important patient-specific modification is to change the type from RDF / XML to OWL / XML in order to create object properties has_lifestyle_participant and has_breakfast_meal. RDF / XML is a serialization syntax for RDF graphics. OWL / XML is a serialization syntax for the OWL 2 structural specification. RDF / XML ontologies cannot be represented correctly using standard XML tools. In addition, there was a desire for a more regular and more extensive XML format. That's why OWL / XML was invented.

Another significant modification is related to the expansion of the "patient profile" to have more than one laboratory test. The DMTO is built as a set of modules. These modules were implemented from scratch or imported from other well-known ontologies. To achieve a personalized diet, different proportions between fats, carbohydrates and proteins are set as dietary parameters for each meal (currently only for breakfast). To determine whether a patient has laboratory test results within the normal range or outside the normal range, a set of rules is established that verify this and also determine the required ratio of fats, proteins and carbohydrates according to the results of laboratory tests. The ratio between fats, carbohydrates and proteins is determined by the patient's diet.

Prediction of antimicrobial resistance in metagenomic data.

Predicting the origin of the samples. Large-scale metagenomic studies [128] - [131] are part of a global initiative to study and understand microbiome diversity. High-throughput screening, such as sequences of entire genomes, identifies genetic information to more detailed levels such as species level and can further detect an abundance of eukaryotes, fungi, and viruses. Most methods for analyzing metagenomic sequence data are based on controlled machine learning techniques [132], [133]. Most of these models are limited to predicting site samples.
The classification of samples by origin is usually performed by controlled machine learning methods, which include the division of samples into training and testing sets. In the present work, a preliminary review of some of the well-known methods was made, after which it was decided to focus the study on three of them, which do not include many parameters and are easier to work with, but sufficiently informative within the R-project. In particular, Gradient Boosting Machine (GBM) [137], Random forest [138] and Neural network (NNet) [139] were used. The applied machine learning models were used to predict which continent and which city the samples belonged to.
Relative risk assessment using spatial modeling. Spatial autocorrelation is very often used when observations that are close in space have similar values. Part of this spatial autocorrelation can be modeled by known covariant risk factors in a regression model, but it is common for the spatial structure to remain in the residues after taking into account these covariant effects. Then spatial models such as Bayesian hierarchical models are used to extend the linear predictor with a set of spatially autocorrelated random effects depending on the neighborhood structure of the geographical areas. Random effects are usually represented by conditional autoregression (CAR) [140], which causes spatial autocorrelation through the neighborhood structure of areal units. Such models are commonly used in epidemiology, such as disease mapping studies [141], but are relatively new in the field of metagenomics.

Omics data compression

When integrating biomedical data, a key issue is the way high-volume data is transferred. This is especially important in the case of so-called omix data and any other data of biomedical origin, which as stated above in the thesis are heterogeneous in type, format, origin and time of generation. Particular priority in the science of biomedical data has been given to the way they are compressed, both as a method and as a software solution.
The proposed approach for compressing biomedical omix data in is based on coding and more specifically on the Shannon-Fano and Huffman algorithms. The main idea behind the Shannon-Fano and Huffman algorithms is to set a binary code that corresponds to each of the characters in the input message (in this case, the bases of the input sequence).
Create optimal code. An optimized algorithm developed in the study, according to which the sequences based on the theory of optimal letter coding will be compressed, has the following operational characteristics:
1. Sequence type recognition (RNA or DNA).
2. Depending on the sequence, predefined binary values are selected for each of the symbols used in the sequence - the use of predefined values preserves the following procedures necessary to build an optimal code for each random message in the Shannon-Fano and Huffman algorithms: a) The set of symbols of source A is arranged in order to reduce the probability of occurrence of a message - pj b) The set of probabilities Pi is divided into two groups (p1, p2, ..., pj) and (pj + 1, pj + 2, ..., pm), so that the difference shown below is minimal: c) The symbols whose probabilities are in the first group are assigned with the rth code letter 0, and those in the second group - with r Code letter 1. d) For a one-element probability group the procedure is completed and for each of the other groups in the multi-stage procedure the elements are numbered from 1 to m and go recursively (r = r + 1) to step 2.
This approach eliminates the shortcomings of both Shannon-Fano and Huffman algorithms for decoding and protecting information. In the proposed optimized approach, each of the symbols

will be encoded with exactly 2 bits. As before compression, each character took up 8 bits, here is the first place where we have a 1: 4 compression. After applying the algorithms, compression of up to 1: 400 can be achieved.

Chapter 4: Software implementation of intelligent systems for integration, analysis and classification of biomedical data

Designing architecture to meet the hardware and software requirements of the system. All system solutions in this paper are based on the same intelligent structure for hardware and software. The heterogeneous data integration system is based on the distribution of data in multiple databases distributed across multiple hardware devices. Two database management systems MongoDB version 4.4 and Neo4j version 4.2 were used. The driver provided by the official distributor is used to connect to each of the databases. All realized software products are installed on an author's developed platform based on Kubernetes [C1].
Figure 4.2 shows a system that allows you to run and maintain applications for which an image of a container (image) has been created. Accordingly, the necessary images of all services used in the thesis are constructed. They are built for MongoDB, Neo4j, RabbitMQ ,, Go based applications, Python based applications and Java based applications. Each of the applications must work in at least 2 instances. This allows even one of the servers to stop working due to external factors such as loss of connection, burned data media or other reasons for the application to continue working. A special controller has been created in Kubernetes, which allows as soon as one of the containers is stopped for some reason not to try to re-create it on the same server as default, but to create it on another server. This allows the application to return to normal operation much faster. When using container-based databases, synchronization between them and replication is very important so that if a container stops, the database can continue to work.

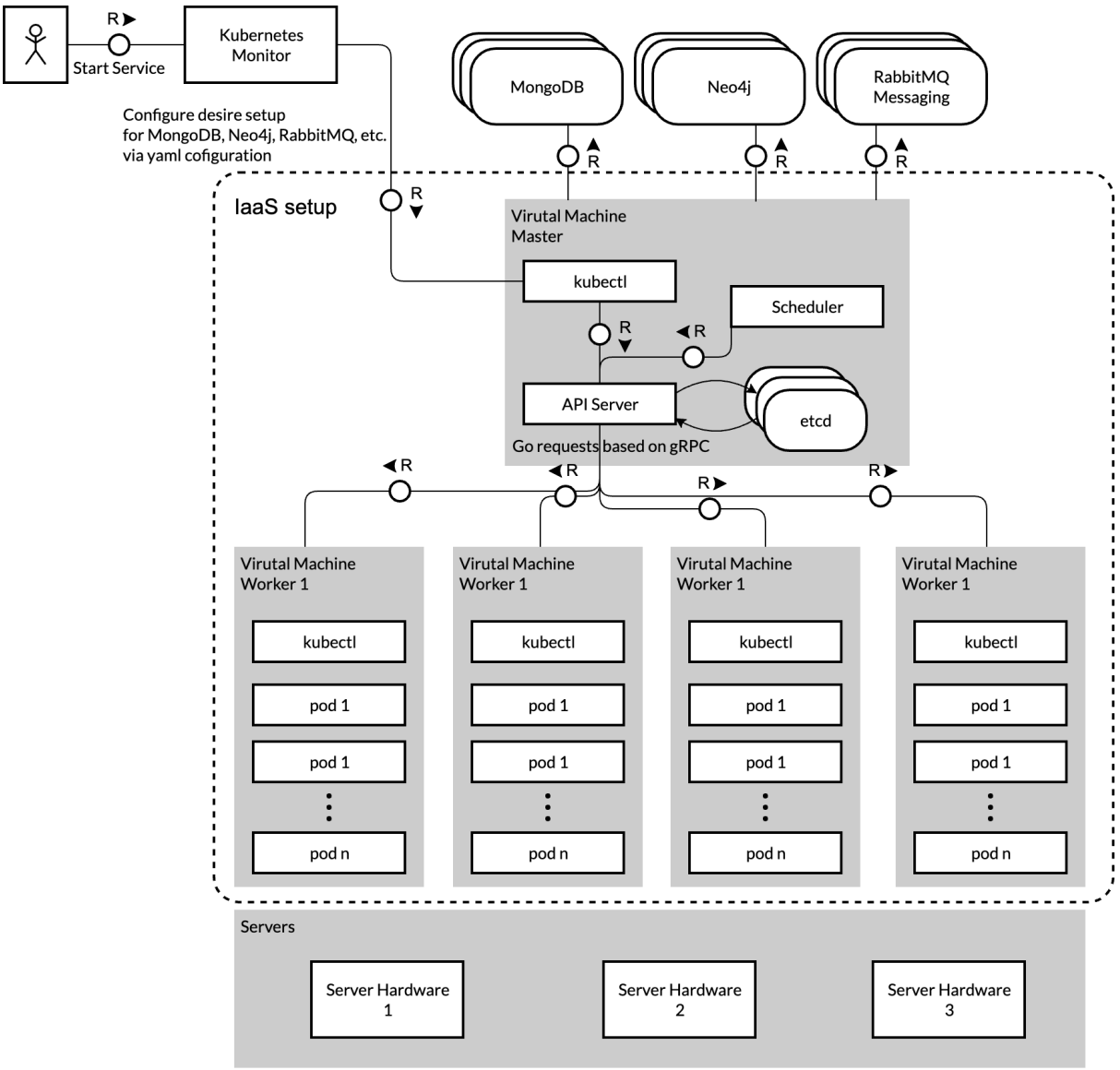


Figure 4.2. Architecture of the developed system for providing hardware resources using three hardware servers.

Heterogeneous data integration and machine learning software module for predicting the survival of cancer patients

Using the semi-structure, heterogeneous data is integrated into a single database, where the ultimate goal is to create a network of connections between all types of data. In this network, nodes represent patients and edges represent similarities between patient profiles. The similarity means that two patients are linked to each other by multiple proteins, based on expression profiles and changes in the number of copies. These networks of connections facilitate the grouping of patients. Patient groups may then be associated with different clinical outcomes.

The network has two layers. The first layer covers the internal relationships built with raw data, ie. clinical information, expression data and copy number variants. They are transformed into patient-protein relationships. The second layer includes semantically related data from sources of knowledge from external domains. These sources provide information on additional proteins related to those existing in our dataset. These new relationships are stored in our chart-based database. In order to use additional information from external sources of

knowledge, they are connected to a network via hyperlinks (URLs). This avoids visual incomprehensibility, which would be caused by an excess of information. These two layers are combined in one network, where each connection is weighed.

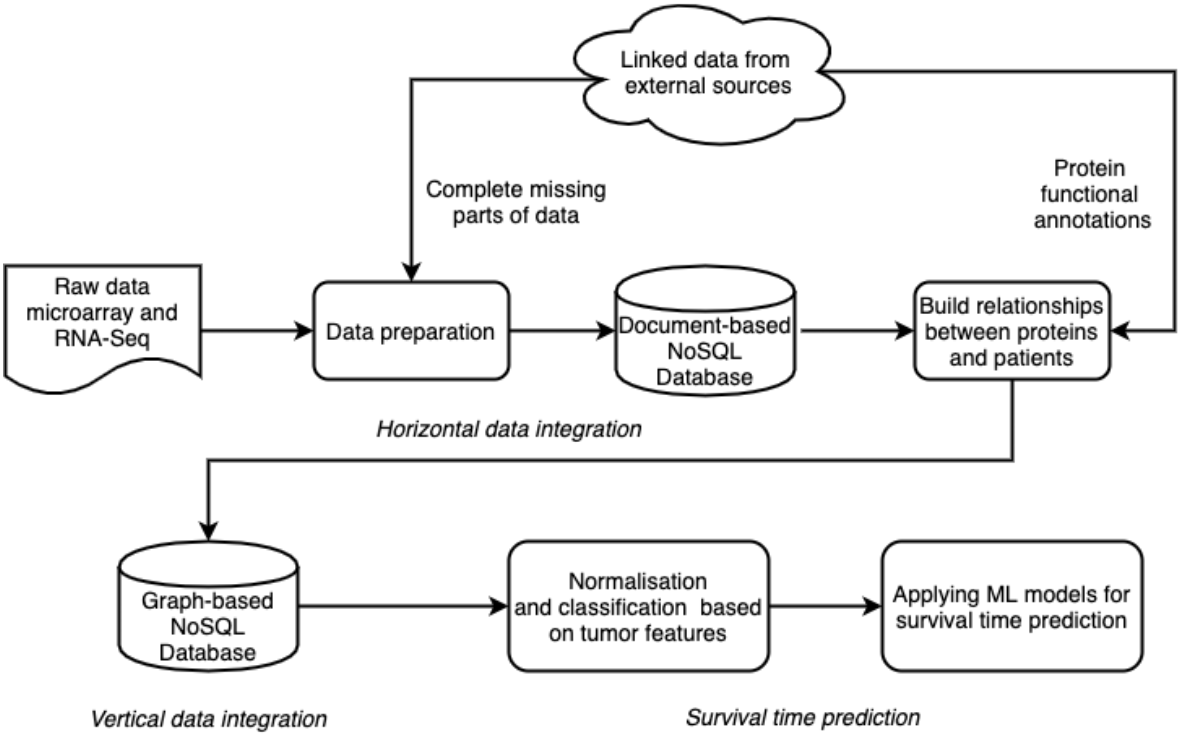


Figure 4.3. Workflow of integration and analysis of cancer data

All data from the experimental datasets are integrated horizontally with NoSQL technology (MongoDB) and are presented as a semi-structure. This results in a semi-structure for each data type, ie. all clinical data are combined into a semi-structure, all expression data into another semi-structure and all copy number data (CNV) into a semi-structure. All data and metadata are stored in MongoDB in JSON format. For the vertical integration of the data, we must first find connections between the already established semi-structures for clinical records, expressive profiles and data for copy numbers. These relationships are managed in the graph-based database - Neo4j. For example, patient A with semi-structure {ID, [attributes]} is associated with patient B with semi-structure {ID, [attributes]}. In this regard, the identifier is the important key, while the attributes provide general information about the type of data record (clinical, expression, copy number). Such relationships facilitate the construction of an individual network for each patient examined. This network includes expression profiles, copy number and mutated proteins. In this way, we can detect and connect all patients through a specific set of expressed and mutated proteins.

Use remote sources for connected data. By integrating semantic data, in particular through https RESTFul endpoints (programming of access points), we can find additional links between proteins from external Domain Knowledge Sources (EDKS), such as gene ontology GO), UniProt, Ensemble. Proteins that are closely related to those present in expression profiles can be detected by EDKS.

A new and universal prognostic parameter has been developed - integrated clinical characteristics of the tumor (TICF). To predict the patient's survival time (in both cancer studies taken together), specific informative clinical features were selected. Different functions and their combinations and order were tested, and the optimal setting shown in Figure 3.8 was empirically established. In Chapter 3. In particular, the TICF characteristic is constructed by numerically combining the tumor stage, tumor size and age at diagnosis (Fig. 3.8) in this exact order. The order of aggregation of clinical data also shows the importance of clinical information for tumor development and the importance of patient survival.

Predicting the survival of cancer patients. The machine learning models used in the developed approach are based on Support Vector Regression (SVR) with different cores: Radial Basis Function (RBF), Linear and Poly (nomial), and Decision Tree Regression - DTR). Such models have been shown to perform well in predicting survival in cancer studies [144], [145]. In addition, the use of these models helps to seamlessly cross-validate the results.

Results of the software implementation of the models. A new network-based data integration model has been developed, combining clinical and molecular data using both raw data records and external sources of knowledge. The links obtained from the raw data represent the internal network, and the links based on external domain knowledge sources (EDKS) are presented as a semantically connected network. The built semantically connected network is connected to EDKS through access endpoints based on RESTFul API / s. Once the data set is normalized and patients are divided into groups, several machine learning models are applied to predict survival time: Vector regression support (SVR with RBF, linear and polynomial cores), and decision tree regression (DTR).

Survival time was predicted using data from both cancers combined with the model used to process and integrate the data. DTR and SVR-Linear perform best, with SVR-Linear giving the most accurate results for predicting survival time. The potential of these models is in improving the accuracy of survival time forecasting by iteratively improving the training data set across the integrated data set. In particular, with each new patient studied for whom the model is used, the data set for training is actually enriched with new trusted relationships determined by the increased frequency of their use.

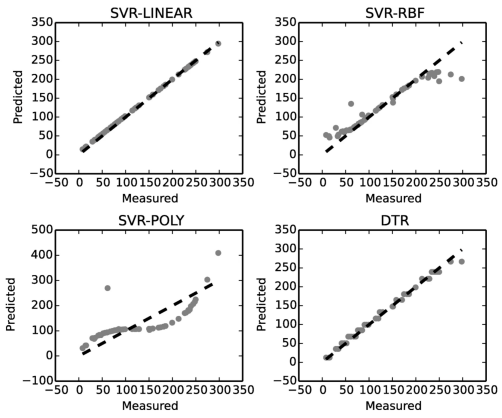
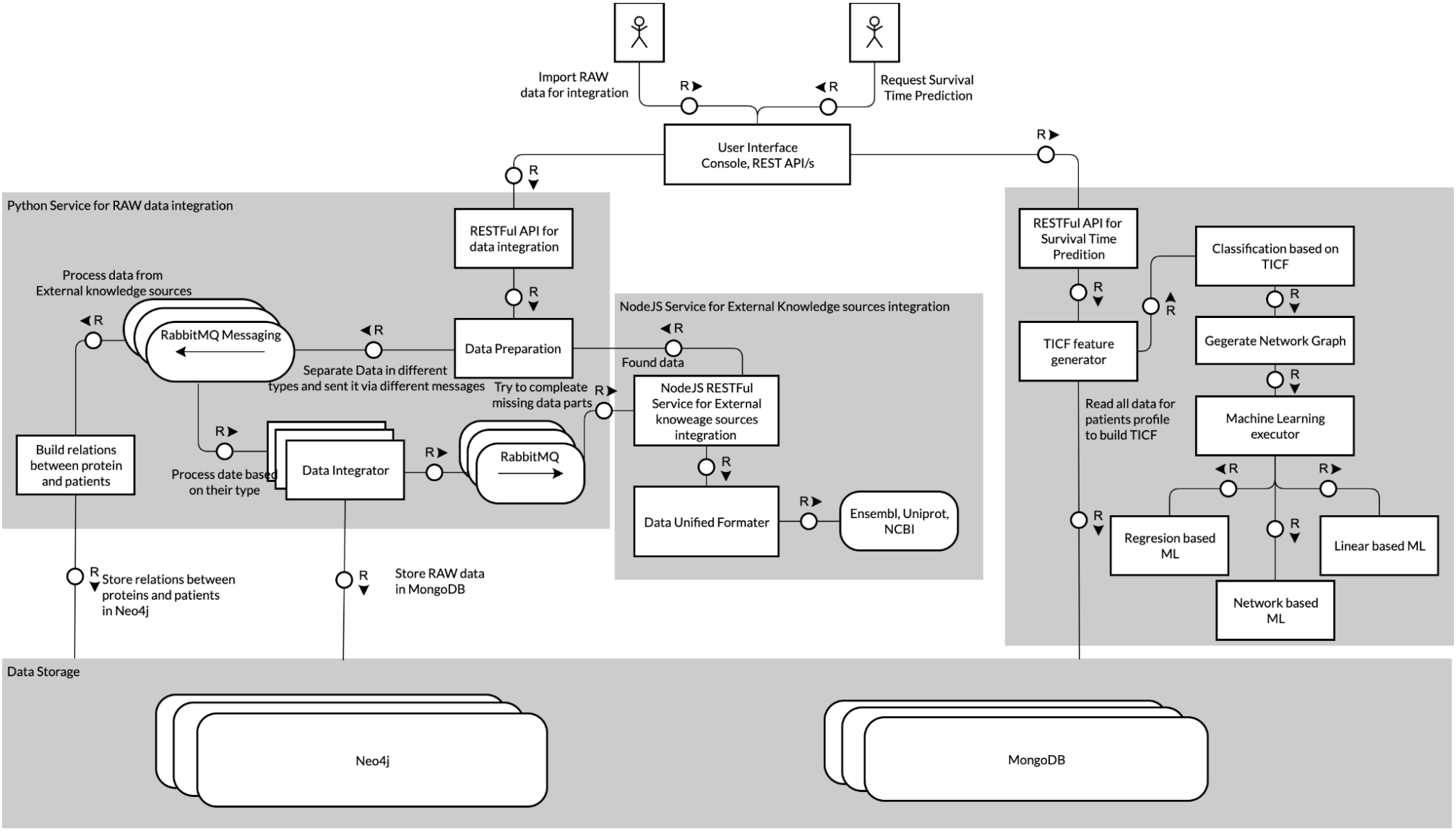


Figure 4.5. Degree of success of machine learning models for survival time prediction.

The predicted and measured values correspond to the TICF of the forecast in the months for survival. The dotted line symbolizes the ideal case of the predicted ratio to the measured TICF for predicting survival time.
The performance of three models (SVR-Linear, SVR-RBF and DTR) is compared. The fourth model, SVR-Poly, showed worse results (Fig. 4.6). Cross-validation is based on 5 subgroups defined by the TICF property obtained after using the k-fold algorithm.

Software for the implemented system. The system is implemented through the built infrastructure presented in section 4.1. The following services have been developed on this infrastructure. Two MongoDB and Neo4j databases, as well as multiple Python, NodeJS and RabbitMQ services (Figure 4.7.)



Фигура 4.7.Софтуерна архитектура на изградената система

The raw data pre-processing and integration module is built on Python version 3.7. Numerous software modules for data integration with different JSON, CSV and XML formats have been implemented. These are the formats with which most publicly available data resources in the bioinformatics world work, such as NCBI, Ensembl, UniProt and others. The integration is realized on the principle of an infinite series of lines that must be described and analyzed. The user submits files via the so-called HTTP 2.0 protocol "stream", which allows a file or resource to be sent in parts. This in turn allows the server part to start processing the file without sending it in its entirety. On the other hand, in bioinformatics, sequential data files have a fairly large volume of over 100 megabytes per file. Without using such an approach to split the file stream into small processing parts, it would limit the maximum number of parallel requests that can be processed at the same time.

Software solutions and results for protein structure prediction and accuracy assessment.

The methodology of this study includes several stages, such as integration of the necessary data from EKDS, data reprocessing, analysis and classification of data based on machine learning. After reprocessing and normalizing the data set, three regression ML models are applied: lasso regression, nearest neighbor regression, and decision tree regression. We test the strategy for cross-validation of the division of k = 3, k = 5 and k = 7 times to compare the models in terms of their accuracy in predicting KB energy results and probability. The graphs in fig. 4.9, 4.10 and 4.11 show the relationship of the actual with the predicted values of each specific model used for KB energy and respectively for the probability of cross-validation (cv) k = 5.

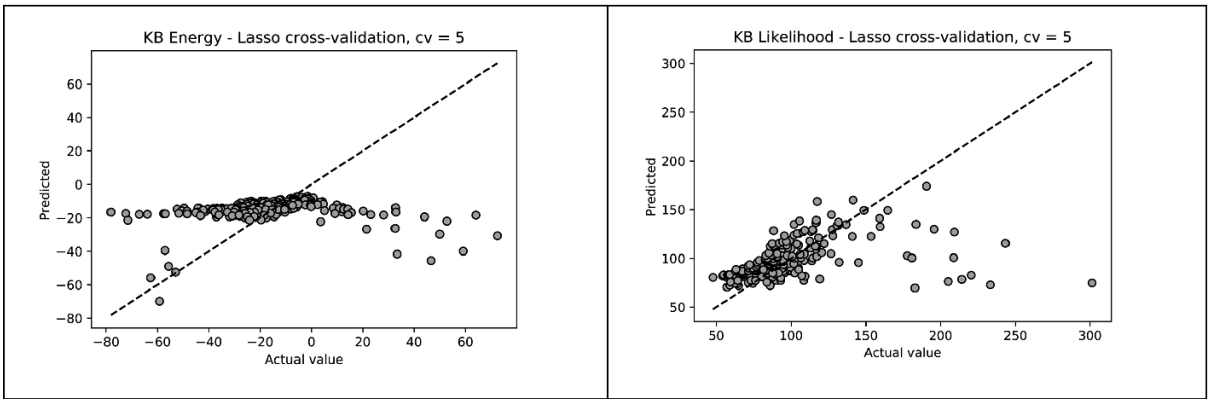


Figure 4.9. Using the LASSO machine learning model for validation with a value of the coefficient k = 5

In terms of knowledge-based energy (KB), the lasso has worse predictive results than the probability of distributing the results around the regression line, with very few deviations from the higher actual probability value. The NNR results are similar, with KB energy estimates being more scattered than probability values. DTR gives somewhat similar results for predicting KB energy and probability values. These results are evidence that the use of the probability forecast is better than the KB energy forecast. These results confirm the analytical computational approach of the optimization finding that the probability is better than the energy of KB as a scoring function.

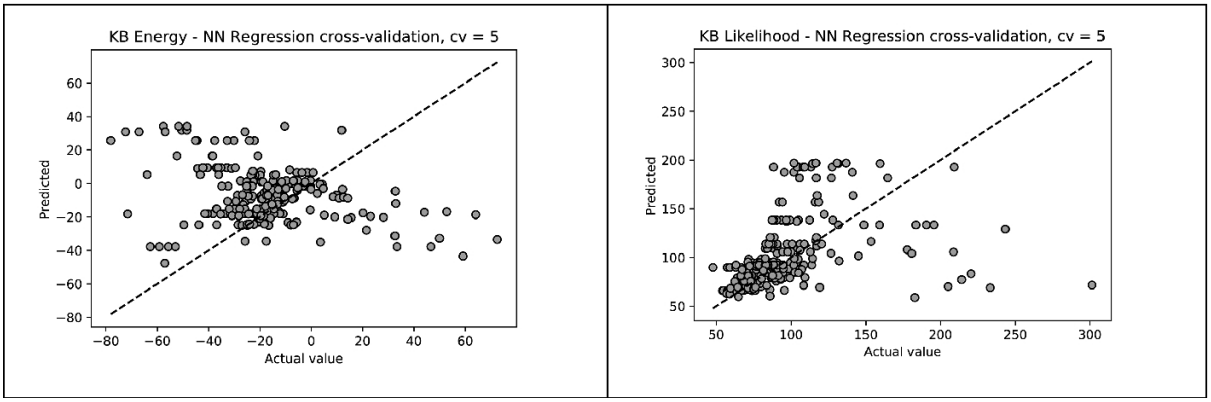


Figure 4.10. Using the model for machine learning "Nearest neighbor regression" for validation with a value of the coefficient k = 5

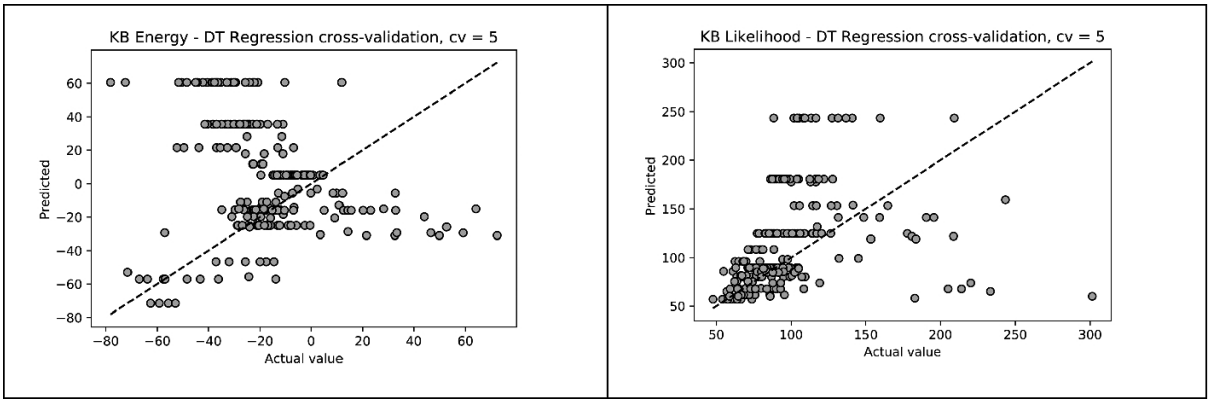


Figure 4.11. Using the model for machine learning "Decision tree regression" for validation with a value of the coefficient k = 5

Software solutions and results in order to create a counseling system for diet counseling in patients with diabetes.

System Architecture: The architectural plan and the general functional plan of the developed DSS are presented in fig. 4.15. The proposed system (Fig. 4.15) consists of the following parts: the input part (patient data, food data), the user integration point with the RESTFul API service, the subsequent data integrator, knowledge base and the corresponding output mechanism and part for storage, including: cache for dietary recommendations, food data storage and patient database. The user integration endpoint server is designed for the purposes of data integration, data normalization, and interface development and implementation. The input is based on patient data and some food data needed to generate dietary recommendations. The model of patient data is determined by the indicators of outpatient tests and medical history (history) as part of the patient's health card. These patient data include all information from laboratory tests and its main components such as glycated hemoglobin, glucose, cholesterol, uric acid. Patients were created from the patient data and recorded in the DMTO.

The main component of the developed DSS for dietary recommendations is the knowledge base for the subject area, the core of which is DMTO. Specifically, the DSS knowledge base consists of two main parts - an extensible copy of the DMTO and a set of SWRL rules describing specific knowledge for data analysis and decision making. The system has developed an application interface as server endpoints based on the RESTFul API, allowing the user to add patient data. Each added set of outpatient records is associated with a specific patient. For each new patient, his identifiers (GUID) and profile are generated. Outpatient tests are linked to the patient's profile through the has_lab_test property of the respective patient instance. When

outpatient records are added to the DMTO, the changes are saved and the generated identifiers of all new patient instances are returned.

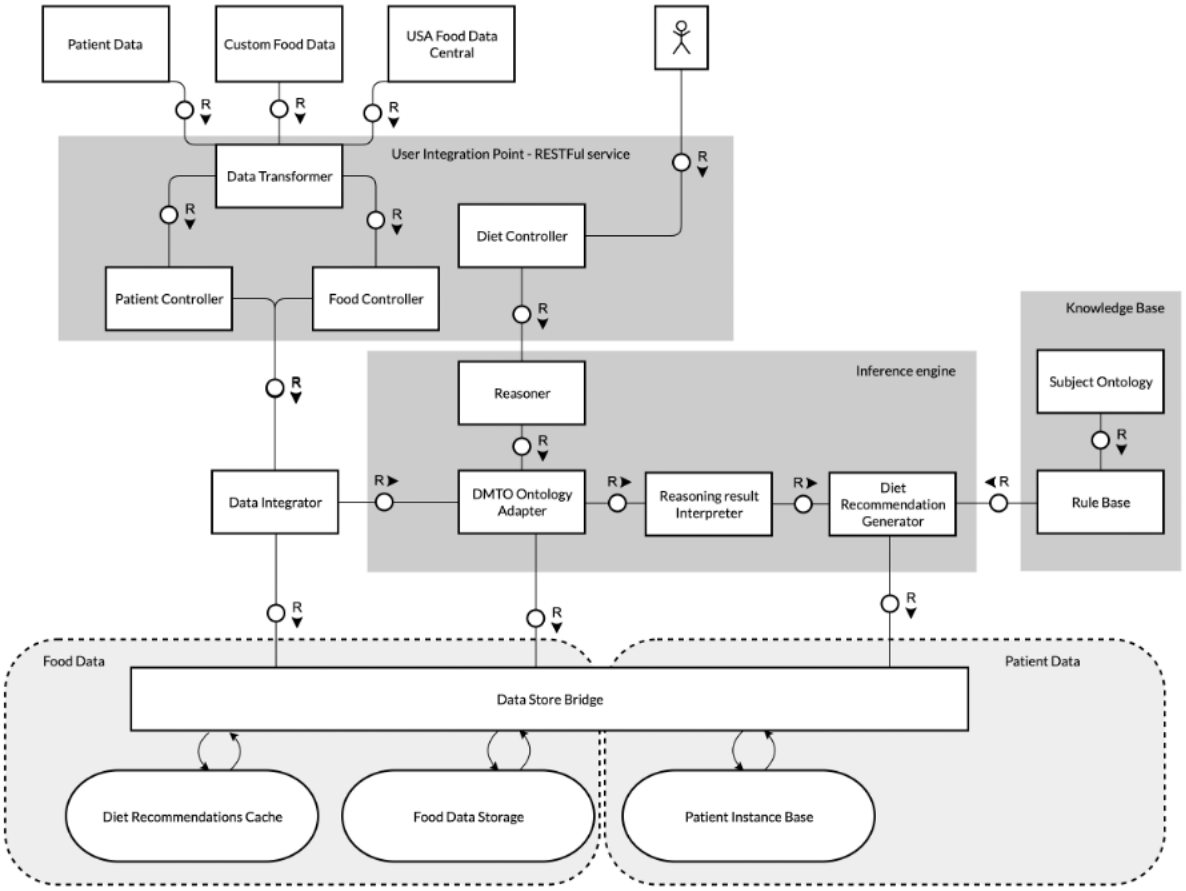


Figure 4.15. Architecture of the developed system

DMTO changes (enrichment). The DMTO is a comprehensive ontology and provides the largest coverage and most complete picture of coded knowledge of the current status of T2DM patients, previous profiles, and T2DM-related aspects, including symptoms, outpatient tests, complications, glucose-related interactions, and medications. , and treatment plan frameworks. The main contribution of the present study is the appropriate extension of the DMTO and the development of a DSS model for dietary recommendations on aspects of T2DM and to support the development of potential post-clinical treatment plans.

Implemented model and software solution for integration, classification and analysis of metagenomic data

The results of this part of the thesis are the result of work on the MetaSUB and CAMDA projects. As noted above in the paper, the aim of these projects is to study microbial, bacterial and viral diversity, as well as to classify and assess the antimicrobial resistance of the studied societies by microorganisms. The data are from metagenomic parallel (NGS) sequencing and are samples from publicly accessible locations, from the transport system of different cities around the world, where diversity is assumed to be high and the level of antimicrobial resistance is significant.

The antimicrobial resistance genes and the corresponding bacterial taxa represent a relatively small part of the available global metagenomic profile. Based on the Kaiju metagenomic classifier, which uses a modified search using efficient Burrows-Wheeler transformation memory [151], it was found that the relative abundance of antimicrobial-related species averaged between 0 and 0.33 of the total sequence data. . Some cities show greater

diversity and numbers such as Fairbanks (max 0.28), Lisbon (max 0.2), Ilorin (max 0.33). The most common antimicrobial taxa are shown in Figures 4.17. One of the clusters includes *Salmonella enterica, Staphylococcus aureus and Escherichia coli,* which are widespread in Ofa, Ilorin and Lisbon. Antimicrobial genes in Streptomyces-related classes are more common in major cities such as London, New York, Hong Kong and Kuala Lumpur. Samples from Berlin, Tokyo, Stockholm and Doha have small or zero numbers among the richest antimicrobial taxa.

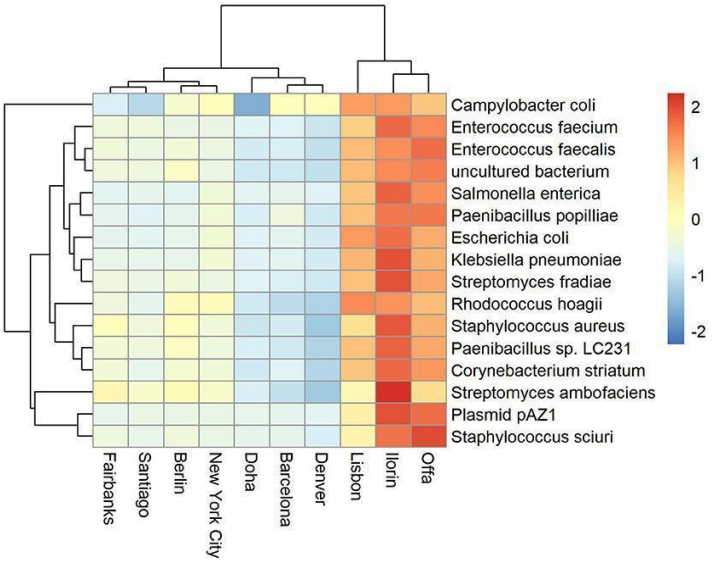


Figure 4.17. Distribution of antimicrobial resistant taxa

Some values of correlations associated with antimicrobial taxa (correlation> 0.6, p <0.01) with meteorological data in cities are: different measures of humidity variability and *Vibrio parahaemolyticus; medium humidity and Campylobacter jejuni, Corynebacterium striatum, Paenibacillus sp*. LC231, *Rhodococcus hoagii, Streptococcus australis and Streptomyces ambofaciens; temperature and Pseudomonas aeruginosa. Vibrio parahaemolyticus and P. aeruginosa* show the strongest negative correlations with humidity and pressure variability, respectively.

Predicting the Origin of Samples. To predict the origin of the sample, three common machine learning techniques are used: Gradient Enhancement Machine (GBM), Random Forest (RF), and Neural Network (NNet). Recursive Feature Removal (RFE) is applied to select the best features for the models. This external sampling method is based on 10-fold cross-validation and is repeated 3 times. The k-fold approach involves dividing the set of data into k groups of approximately equal size. The first set is treated as a validation set and the method is to validate the other k-1 groups, where k is generally assumed to be 5 or 10.

Spatial analysis. For spatial analysis, all available genes are used to make a convolutional model. The spatial correlation in the cities was made by the Moran I-test. Cities such as New York (max. 0.44, p <0.01), Ilorin (0.38, p <0.01), Hong Kong (0.41, p <0.01) and Taipei (0.6 , p <0.01) show strong spatial correlations for many antimicrobial resistant taxa. Data on the number of metagenomics often show over-variance, as they are heterogeneous due to different cities and countries. An over-dispersion test was performed for the 16 antimicrobial characteristics of the prediction models, combining a Poisson model with covariates and a simple least-square regression to estimate the over-dispersion parameter. The results show that all but one of the best antimicrobial characteristics appear overdispersion with p-values well below 0.01.

Discussion of the results and software implementation. This paper demonstrates the three machine learning methods, namely Gradient Boosting machine, Random Forest and Neural Network, which have similar performance for classifying the origin of samples. Using a large database such as proGenomes, which contains over 80,000 annotated bacterial and archaeological genomes, we achieve high accuracy (up to 80%). The developed program modules generate all tables and figures so that the results can be reproduced. In addition, the code allows users to change parameters, for example by using a different set of settings and also to perform additional machine learning methods as provided within the package, and to further improve the results.

Implemented model and software solution for sequential data compression

A new model and software solution for data compression from parallel (NGS) sequencing has been implemented. The methodology and architecture of the proposed model is explained in detail in section three. C ++ was chosen for the implementation of letter and noise coding algorithms because of its execution speed and lower memory usage compared to other languages. It also allows direct access to memory, easy and intuitive implementation of bitwise operations and low-level resource management. These advantages make algorithms written in C ++ more optimal than high-level languages such as C # and Java. This is one of the main reasons for including cross-platform links between C ++ and high-level languages.

The combination of C ++ (unmanaged code) with C # (managed code) is realized thanks to the .NET platform. Even at its creation, the idea of independence from the environment was embedded. The source code is not compiled to instructions designed for a specific microprocessor and does not use specific capabilities of a particular operating system, but is compiled to an intermediate language - Common Intermediate Language (CIL). This language is not executed directly by the microprocessor, but by a virtual environment for executing CIL code, called - Common Language Runtime (CLR)

User interface. The user interface architecture is based on the vue.js framework. It provides many components for the implementation of the interface, each of which is built on a modular principle and can be easily expanded. The entry point is through a browser, and the tested ones are Safari, Mozila, Opera, Edge, Google Chrome and others. This is possible because each of the components is written in JavaScript according to the ECMA 6 standard. The generated code is then compiled to HTML depending on the browser that executed the request. The V8 software framework developed by Google is used to interpret and compile JavaScript to an abstract syntax tree from where it can be compiled as C ++, C, C #, Java and others. Node.js is used as the software environment for running Javascript. In this way, the user interface is made up of a server and a user clock. The constructed components work on the realized infrastructure in point 4.1.

Architecture of the built system for compression of sequential data. All components described above are combined in one system presented in Figure 4.27. The system is divided into four main modules. Module with the implementation of all supported coding methods including the modified Huffman method. This modified method has only one difference from the original and that is that a preset table is selected for the alphabet of the code. This speeds up the coding process many times over and also avoids the possibility of choosing the wrong alphabet, which leads to an increase in the degree of compression. The developed algorithms are entirely based on C ++. All algorithms have a common abstraction, which allows very easily to expand the set of compression algorithms. They have no external dependencies on system software resources for working with memory or using the operating system. Only the standard modules built into C ++ are used.

A C ++ module for HTTP-based communication via RESTFul has been built. This makes it easier to use the library. There is a separate entry point for each of the algorithms, which makes it easier to work with them and allows them to be used independently of each other. The HTTP 2.0 protocol is used, which allows files to be transmitted in fragments. Due to the way coding algorithms work, only Huffman code optimization is able to compress in real time. This is possible because it does not require a frequency analysis to determine the table of the most common symbols. This is a huge advantage that achieves real-time compression of any size DNA / RNA sequence.

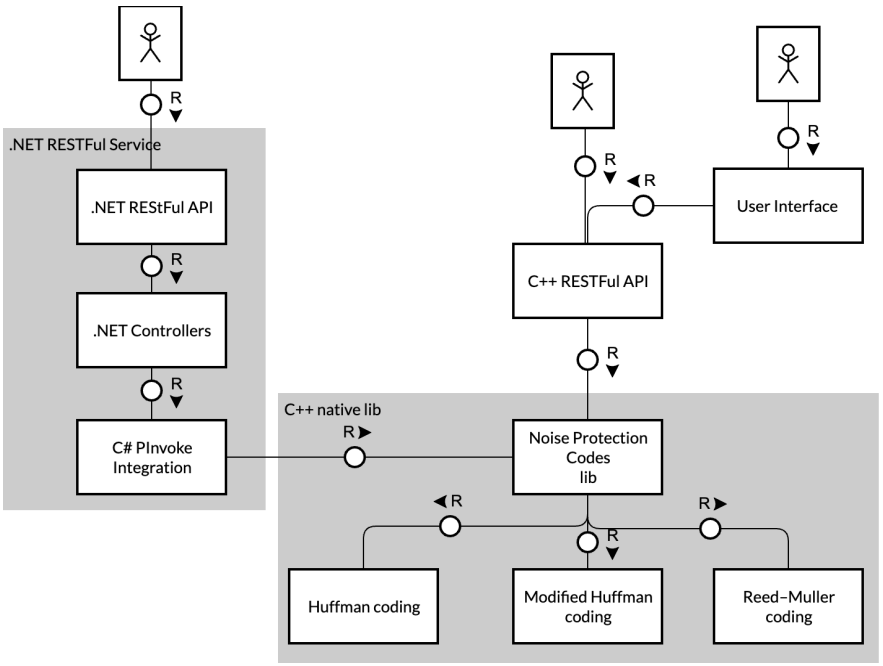


Figure 4.27. Architecture of the built system for compression of sequential data

Chapter 5: Contributions and Perspectives

A set of tools for intelligent semantic integration, classification and analysis of biomedical data through the use of computer science tools and artificial intelligence is proposed. Many author's information systems have been built on the basis of an author's platform. The qualities of the proposed tools are studied in detail using validation methods.

Theoretical and methodological contributions of the thesis

1. A model and software implementation for integration of heterogeneous data using intelligent systems has been developed. (3.1, 3.2).

2. A model and software implementation for semantic integration of data from different cancers have been developed (3.3).

3. A model and software implementation have been developed to predict the survival of cancer patients using informatics and artificial intelligence tools (3.4).

4. A model and software implementation have been developed for extracting knowledge from semantically integrated data using an example of a dietary counseling system for patients with diabetes (3.5).

5. A model and software implementation for prediction of antimicrobial resistance in metagenomic data have been developed. Used to classify samples from different countries and continents (3.6).

6. A model and software implementation for sequential data compression using noise protection coding algorithms (3.7) have been developed.

7. A platform for the provision of software as a service has been developed, which is used for the implementation of all systems in the thesis (4.1).

Experimental and practical contributions to the thesis

1. Experimental confirmation of the benefit of horizontal and vertical data integration (4.2) has been achieved.
2. Experimental confirmation of the benefit of applying machine learning in predicting the survival of cancer patients and in data validation has been achieved (4.2).
3. Numerous machine learning models have been developed and successfully implemented in the context of predicting the survival of cancer patients (3.3.4).
4. Machine learning models have been used and practical results have been achieved regarding the prediction of protein structures, their classification and accuracy assessment analysis (3.4).
5. In-depth work has been done to integrate, classify and analyze data from metagenomic studies on prevalence, genetic diversity and antimicrobial resistance.
6. Experimental results have been achieved confirming the need to use an advisory system based on ontologies for offering diets in accordance with a diet in patients with type 2 diabetes (3.5.1).
7. Experimental results have been achieved confirming the need for sequential data compression and the need to create a new compression model that can work with the new technological paradigms (4.6).

List of author's contributions

1. A model and the corresponding software implementation for intelligent horizontal and vertical integration of biomedical data are proposed (3.1.3).
2. Created on an architecture and model for semantic integration of data for both cancer and diabetes patients (3.2.1, 3.3.2).
3. Created a new complex characteristic for classification and prediction of survival of cancer patients (3.3.4).
4. Established system for prediction and validation of protein structures (3.4)
5. Established a dietary advisory system for patients with diabetes (4.4).
6. A system for predicting the survival of cancer patients has been established (4.2.1).
7. A model and software implementation for classification and spatial modeling of sources and antimicrobial resistance in metagenomic data have been created (4.5.2, 4.5.3.).
8. A model and system for sequential data compression has been created, allowing real-time compression (4.6.2).

Prospects for future development

The methods presented in this paper for semantic integration, analysis and classification of biomedical data through the creation of intelligent systems is a good basis for future work in which to cover a wider range of tasks, as well as to enter into more in-depth computational research.

1. The considered methods for horizontal and vertical data integration can potentially be applied with other types of data than biomedical informatics.
2. The semantic integration methods considered can be extended to use additional ontologies in order to create better semantic connectivity between the data.
3. The considered models for predicting and validating the survival of cancer patients and protein structures can be expanded with new machine learning methods.
4. The extension of the established dietary counseling system for patients with type 2 diabetes can be improved through additional knowledge from medical centers, as well as

the introduction of more features to be considered when creating a diet. Such a system should be integrated into the development of the country's electronic health record.

5. The established models for classification and spatial modeling of antimicrobial data can be extended by adding databases for existing cities, as well as adding additional validation to increase the accuracy of the models.
6. The established sequence data compression system can be extended by adding a module that allows integration with known sequence data storage systems such as NCBI.

Declaration of originality

In line with the procedure for obtaining the educational and scientific degree "PhD" at Sofia University "St. Kliment Ohridski" based on the defense of the thesis presented by me, I declare that:

- The results and contributions to the research, presented in my thesis "Intelligent information systems in bioinformatics: semantic integration, analysis and classification of biomedical data", are original and they are not borrowed from research and publications in which I do not participate as a co-author.
- In the text of my thesis there are not illegally used texts and other objects of copyright without specifying the source, or without the permission or legal right to do so.
- My thesis is not applied for obtaining a scientific degree in another higher school, university or research institute.
- The information presented by me as a list of publications, copies of documents and results obtained during the experiments - are objective truth.

Publications on the topic of the thesis

[C1] H. Sabev, **I. Mihaylov,** and R. Rashidov, “Distributed persistent virtual machine pooling service,” Patent: US10824461B2, Nov. 03, 2020.

[C2] **I. Mihaylov,** M. Kańdula, M. Krachunov, and D. Vassilev, “A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models,” Biology Direct, vol. 14, no. 1, pp. 1–17, 2019.
Импакт фактор = 2.913, SJR = 1.388, Q1, цитирания: **16**

[C3] **Mihaylov, I**., Nisheva, M., Vassilev, D., “Machine Learning Techniques for Survival Time Prediction in Breast Cancer,” Lecture Notes in Computer Science. Lecture in Artificial Intelligence 11089. Springer International Publishing, 2018, pp. 186–194, doi: 10.1007/978-3-319-99344-7_17.
Импакт фактор = 1.14, SJR = 0.43, Q2, цитирания: **5**

[C4] **I. Mihaylov,** M. Nisheva, and D. Vassilev, “Application of machine learning models for survival prognosis in breast cancer studies,” Information, MDPI, vol. 10, no. 3, p. 93, 2019.
SJR = 0.35, Q3, цитирания: **13**

[C5] K. Serafimova, **I. Mihaylov,** D. Vassilev, I. Avdjieva, P. Zielenkiewicz, and S. Kaczanowski, “Using Machine Learning in Accuracy Assessment of Knowledge-Based Energy and Frequency Base Likelihood in Protein Structures,” in Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 572–584, doi: 10.1007/978-3-030-50420-5_43
Импакт фактор = 1.14, SJR = 0.43, Q2, цитирания: **1**

[C6] M. Nisheva-Pavlova, S. Hadzhiyski, **I. Mihaylov,** I. Avdjieva, and D. Vassilev, "Linking Data for Ontology Based Advising in Healthcare," in Proc IEEE Explore 2020 International Conference Automatics and Informatics (ICAI), 2020, pp. 1–5. 10.1109/ICAI50593.2020.9311382

[C7] **I. Mihaylov,** M. Nisheva-Pavlova, and D. Vassilev, "An Approach for Semantic Data Integration in Cancer Studies," in Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 60–73, doi: 10.1007/978-3-030-22744-9_5
Импакт фактор = 1.14, SJR = 0.43, Q2, цитирания: **2**

[C8] Zhelyazkova, M., Yordanova, R., **Mihaylov, I.**, Kirov, S., Tsonev, S., Danko, D., Mason, C., Vassilev, D., "Origin Sample Prediction and Spatial Modeling of Antimicrobial Resistance in Metagenomic Sequencing Data," Front. Genet., vol. 12, 2021, doi: 10.3389/fgene.2021.642991.
Импакт фактор = 3.789, SJR = 1.47, Q1

[C9] Zhelyazkova, M., Yordanova, R., **Mihaylov, I.**, Kirov, S., Tsonev, S., Danko, D., Vassilev, D. Bayesian Hierarchical Modelling for Antimicrobial Resistance Abundance. In Proc. International Symposium on Bioinformatics and Biomedicine, (http://bioinfomed.org) pp 28-29, 8-10, October, Bourgas, Bulgaria.

[C10] B. Pulova-Mihaylova, **I. Mihaylov,** I. Avdjieva, and D. Vassilev, "A System for Compression of Sequencing Data," 2020, ISGT, pp 223-235, http://ceur-ws.org/Vol-2656/paper23.pdf
SJR = 0.18