Breast cancer is the most common pathology among women. The death rate from breast cancer among women remains high. Early diagnosis and individual therapy are effective ways to extend people’s lives. The main diagnostic methods are cytological, histological, and immunohistochemical. The cytological method allows assessing the qualitative and quantitative changes in cells, as well as identifying extra- and intracellular inclusions and microorganisms. The histological method allows you to explore changes in the location of groups of cells in a particular tissue. The immunohistochemical method is based on the use of biomarkers. Immunohistochemical images are the result of an immunohistochemical investigation. The aim of the work is to develop a database of immunohistochemical images of breast cancer. With the developed database, a database design methodology was used, including infological, datalogical and physical design. The scientific novelty lies in the use of an object-oriented approach for designing a database of immunohistochemical images. The practical value of the work lies in the development of all stages of database design. As a result, an infological model, a data model, and a UML database diagram have been developed. For the practical implementation of the server part of the database, operating systems such as Windows/Linux/macos can be used, the database server is MySQL. The developed breast cancer database contains more than 500 images for four diagnoses. The image resolution is 4096 x 3286 pixels. For each image, two features are given: relative area and brightness level. The developed Hi&IHCIDB database has medium volume, high resolution, and quantitative characteristics in the description of immunohistochemical images.

Key words: immunohistochemical images, database, infological model, datalogical model, breast cancer.

Introduction

Breast cancer is the most common cancer among women. That is about 30% of all new cancers in women each year. Breast cancer mostly occurs among middle-aged and older women. The median age at the time of diagnosis of breast cancer is 62 years. A small number of women diagnosed with breast cancer are younger than 45 years of age. The average risk of developing breast cancer in women is about 13%.

Since 2007, the death rate from breast cancer among women under 50 has remained stable but continues to decline in older women. From 2013 to 2018, the death rate was down 1% per year. The decrease in mortality is believed to be the result of early detection of breast cancer through screening and awareness, as well as better treatment [1].
Early diagnosis and individual therapy are effective ways to extend people's lives. The main diagnostic methods are cytological, histological and immunohistochemical studies. The cytological method allows assessing the qualitative and quantitative changes in cells, as well as identifying extra- and intracellular inclusions and microorganisms. The histological method allows you to explore changes in the location of groups of cells in a particular tissue.

Thanks to such, you can find out what type and stage of the pathological process, and what treatment should be selected to achieve the maximum effect.

Immunohistochemical study is used to personalize the treatment of breast cancer. Most of the tasks of diagnosing breast pathology require additional immunohistochemical. In the of the expression of estrogen and progesterone receptors, both quantitative and qualitative indicators are determined.

In our investigations, the Allred method (Quick score) was used, according to which 2 criteria are evaluated: the number of positive cells (PS - proportion score) and color intensity (IS - intensity score). The points obtained are the total score (Total score). According to this score, the subtype of cancer is determined, and the appropriate treatment is selected. Immunohistochemical samples are used to form images corresponding to the tumor response to a specific receptor (Estrogen, Progesterone, Her2neu, Ki-67). The image data is written to the database. The medical report of the histopathologist is also saved with the images. Therefore, the actual problem is the formation of a database of immunohistochemical inventions used for the correct diagnosis of breast cancer.

Related works

Let us analyze the articles related to the research topic. The research [2] presents CHISEL (Computer-assisted Histopathological Image Segmentation and Evaluation), system that quantifies digitized samples of benign and malignant (breast cancer) tissues with immunohistochemical nuclear staining of varying intensity and varying compactness. The system seamlessly segments images based on region cropping and explicit splitting of kernel clusters with edge refinement.

The paper investigates the early diagnosis of pneumonia based on deep learning methods. The authors propose the approach to detecting mild pneumonia on x-rays and interpreting classification results using CNN. Biomarkers are identified by genes that mutate between normal and cancerous tissues or change the level of protein or RNA expression. Paper [4] defines biomarkers to develop an automated pipeline to compare the subcellular arrangement of proteins between two sets of immunohistochemical images.

The study [5] proposes a heuristic architecture search for medical image classification problems. Implemented an original approach called morphism of the network to the search algorithm. The results of the research [6] reveal the association of MYCN amplification with suppressed cellular immunity and the potential predictive value of infiltrating CD4 T cell transcripts in pediatric NB. The IHC Profiler plug-in developed in [7] is compatible with ImageJ, which creates a profile for pixel-by-pixel analysis of an IHC digital image and additionally evaluation a score in a four-level system. In the work [8], fully convolutional neural networks (FCNs) were studied, and a modified 3D U-Net architecture was proposed for processing volumetric computed tomography (CT) images in automatic semantic segmentation problems.

The work [9] presents development in the form of a web-based information system with an accessible and comfortable user interface. The information system consists of three main modules. In the article [10], scientists developed a set of tools using a semi-automated scheme suitable for several types of IHC image analysis. It can detect the main components of an IHC image and is useful to researchers for further analysis.

The scheme proposed in work [11] can be easily used for any histopathological diagnostic process that requires a quantitative assessment of nuclear staining and cancer classification. In addition, reduced processing time and manual processing procedure can be implemented in a real-time device to create an online IHC tissue image evaluation system.

Integrin α6 plays a significant role in the onset and progression of lung adenocarcinoma and may act as a prognostic predictor of lung adenocarcinoma in patients. Based on the results of the research [12], α6 integrin may be a potential target gene for the treatment of lung adenocarcinoma.

The paper [13] explores the use of convolutional neural networks for automatic segmentation of immunohistochemical images based on the U-net architecture. The segmentation of breast cancer images and design of intelligent diagnostic systems are considered in the articles [14-16].

The scientists obtained full high-resolution slides (up to 40x zoom) of different types of tissues (impressions, particles of the lungs, mammary gland). The initial size of images varies from 15k x 15k to about 50k x 50k pixels. The resulting images are organized into sets of successive pieces of tissue, where each section was stained with a different dye [17,18].

Analysis of the basic biomedical images databases

As part of the Automatic Non-rigid Histological Image Registration (ANHIR) [19] competition, the database of images stained with various dyes and applied with key points has been published. The competition was part of the "IEEE International Symposium on Biomedical Imaging (ISBI) 2019" symposium.
The project contains a set of images with associated instructions and tools for evaluating proprietary image registration algorithms [20].

The initial stained histological tissue dataset consists of pairs of images of related sections (serial sections). Each image in the pair is colored with a different stain. A minimum of 40 key point landmarks is manually placed on images for evaluation.

Virtual histological laboratory "Histology Guide" is designed to teach visual recognition of the structure of cells and tissues [21]. This site and associated atlas feature the work of T. Clark Brelle and Robert L. Sorenson.

To obtain virtual slides (up to 150,000 x 75,000 pixels), an Aperio slide scanner was used. Site users can interactively explore large images by zooming and panning in real time. Instructions are provided for the images (qualitative description).

Online atlas "Bethesda System" of the American Society of Cytopathology (draft 2014) [22]. For individual diagnoses, one image is shown with interpretation and cytomorphological criteria. For instance:
1. Interpretation: Spindle cell carcinoma.
2. Cytomorphological criteria: Fusiform cells with oval hyperchromatic nuclei, sparse or moderate cytoplasm, nuclear pleomorphism, and nucleoli.
3. Approximately 400 images, 1000x1000 pixels in size.

University of Wisconsin Breast Cancer Diagnostic Data Set (USA). (Not pictured) Breast Cancer Wisconsin (Diagnostic) Data Set (1995) [23]. Features are derived from digitized images of finely ill aspirate (FNA) of breast tissue. Signs reflect the characteristics of cell nuclei. The image is divided by diagnosis: M = malignant, B = benign.

For each cell nucleus, ten features are calculated:
a) radius (average distance from the center to points along the perimeter).  
b) texture (standard deviation of gray values).  
c) perimeter.  
d) area.  
e) smoothness.  
f) compactness.  
g) concavity.  
h) concave points.  
i) symmetry.  
j) fractal dimension.  

A set of annotated images "UCSB Bio-Segmentation Benchmark dataset" for testing computer vision algorithms, in particular segmentation and classification [24].

The data set consists of 2D/3D images and time-lapses that can be used to evaluate the performance of new computer vision algorithms. These images are used for segmentation, classification and tracking tasks (tracking).

For each class of problems, at least one set of reference data is available. Metrics are also given to compare the results of the algorithms. Matlab code is also available to evaluate the performance of algorithms.

The data is organized in two different ways, one based on the type of image content (subcellular, cellular and tissue level data) and the other based on image processing methodology (segmentation, classification or tracking).

The database contains approximately 50 H&E-stained histopathology images used for breast cancer cell detection with corresponding reference data. For such images, an urgent task is cell segmentation for the subsequent classification of benign and malignant cells.

The dataset consists of hematoxylin and eosin (H&E) stained histological images of the breast. The total volume of the database is 400 images, divided by type of cancer [25].

The unresolved problem is the formation of breast cancer database of immunohistochemical images with quantitative characteristics.

The objectives of the study are:
1. Designing the infological database model.  
2. Designing the datalogical database model.  
3. Designing the physical database model.  
4. Conducting experiments and comparing the developed database with known ones.

**Database Structure**

When developing information systems, infological models are often used to represent data. Infological modeling is used to represent information stored in a database. The main elements of the infological model are the entity, attribute, key, connection. The infological model of the developed system is shown in Figure 1.
In this infological model, the following entities are distinguished:

"Attending physician" - this entity describes, in accordance with the name, the attributes of the attending physician, namely: full name, position.

"Patient" is an entity characterized by the presence of the following attributes: full name, gender, date of birth, medical card number.

"Research" - this entity describes the attributes used to store and process information about the investigation. List of attributes: material, operation type, marking, operation date, examination code, examination date.

"Research output" is an entity designed to store information about the results of the study. It has the following attributes: number of slides, staining technique, date of receipt of samples, description of macroscopic examination, description of microscopic examination, description of immunohistochemical examination.

The description of the infological model occurs at the early stages of information systems design. To describe the logical level and display the logical relationships between elements, datalogical design is used. The datalogical model of the database is shown in Figure 2.

Fig. 1. Infological model of the developed system

Fig. 2. Datalogical database model
A detailed model of the database structure for storing information about research results and algorithms used is shown in Figure 3.

To improve the quality of the research, the tables "preprocessing_algorithms" and "segmentation_algorithms" are available. These tables store information about filtering algorithms, brightness adjustment, morphological operations, segmentation algorithms, and their input parameters.

UML class diagram for working with the database in Figure 4. The “storeData” package contains base libraries for working with the MySQL database. The SQL database methods for executing database queries, including connecting to a SQL server, methods for adding, updating, and deleting records. There is also a method for executing SQL queries and a method for obtaining the result of a query to the database.

The “SQLDatabaseParam” class contains the fields and methods required to connect to a MySQL server, such as host, port, user, password, and database name. This class also provides selector methods for setting and retrieving these variables.

Requirements for the server part of the database are shown in table 1.
Requirements for the server part of the database

<table>
<thead>
<tr>
<th>Operating system</th>
<th>Database server</th>
<th>Recommended amount of disk space</th>
<th>Ability to use distributed databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows/Linux/MacOS</td>
<td>MySQL</td>
<td>1 GB</td>
<td>2</td>
</tr>
</tbody>
</table>

Requirements for the storage of immunohistochemical images are given in table 2.

Table 2

<table>
<thead>
<tr>
<th>Operating system</th>
<th>Recommended amount of disk space</th>
<th>Structuring images by catalogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows/Linux/MacOS</td>
<td>100 GB +</td>
<td>+</td>
</tr>
</tbody>
</table>

Experiments

An example of storing information in the table "preprocessing algorithms" is shown in Figure 5.

![Fig. 5. Example of storing information in the table «preprocessing_algorithms»](image)

The "image_characteristics_id" field allows you to store the ID of the input image parameters and the corresponding image pre-processing algorithms, in particular the filtering algorithm, the window size to be applied, the brightness level. In the future, new criteria can be added to improve the quality of immunohistochemical images.

An example of storing information in the table "image_characteristics" is shown in Figure 6.

![Fig. 6. Example of storing information in the table «image characteristics»](image)

The calculation of the relative area and the average brightness level is stored in this database table. The "experiment_images_id" field stores the inputs of the input images, which allows you to further combine the results of research with the image.

Comparisons of databases are shown in table 3.

Table 3

<table>
<thead>
<tr>
<th>Name, type</th>
<th>Organs levels</th>
<th>Amount</th>
<th>Image</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ANHIR H&amp;E, IHC</td>
<td>breast</td>
<td>50 sets</td>
<td>From 15k x 15k to 50k x 50k pixels.</td>
<td>Key points for registration</td>
</tr>
<tr>
<td>2. Histology Guide</td>
<td>-</td>
<td>-</td>
<td>150,000 x 75,000 pixels</td>
<td></td>
</tr>
<tr>
<td>3. Bethesda Cancer H&amp;E</td>
<td>cervical</td>
<td>400</td>
<td>1200x950</td>
<td></td>
</tr>
<tr>
<td>4. Breast Wisconsin H&amp;E</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10 for the cell nucleus</td>
</tr>
<tr>
<td>5. UCSB H&amp;E</td>
<td>Intracellular, cellular, tissue</td>
<td>Over 600</td>
<td>Max 1024x1024</td>
<td></td>
</tr>
<tr>
<td>6. ICIAR2018 H&amp;E</td>
<td>breast</td>
<td>Total 400. (4 diagnoses)</td>
<td>2048 x 1536 pixels</td>
<td></td>
</tr>
<tr>
<td>7. Developed DB (HI&amp;IHCIDB)</td>
<td>breast</td>
<td>Over 500 (4 diagnoses)</td>
<td>4096 x 3286 pixels</td>
<td>2 features for images (relative area and level of brightness)</td>
</tr>
</tbody>
</table>

The developed HI & IHCIDB database has an average volume, high resolution and quantitative features of the description of immunohistochemical images.

Conclusions

The main results of this work are:

1. Based on the analysis of literature sources, the relevance of building a database of immunohistochemical images of breast cancer is shown.
2. Based on the selected entities, attributes, keys and relationships between them, the infological model of the immunohistochemical images database was built.

3. Based on the infological model, a datalogical database model was built, which made it possible to design a physical database model.

4. For the practical implementation of the server part in the database, Windows operating system with MySQL database server was used.

5. The developed HI&IHCIDB breast cancer database contains more than 500 images of four diagnoses. Image resolution is 4096 x 3286 pixels. For each image, two features are given: relative area and brightness level.

The next areas of research are the generation of immunohistochemical images in order to improve the accuracy of image recognition using neural networks.

References


17. Diagnostic Image Analysis Group at Radboud Medical Center. URL: https://ahnir.grand-challenge.org/Data/ (якщо читати: 19.03.2022).


<table>
<thead>
<tr>
<th>Відомий автор</th>
<th>Організація</th>
<th>Електронна адреса</th>
<th>Scopus ID</th>
<th>ResearcherID</th>
<th>Додаткові ресурси</th>
</tr>
</thead>
<tbody>
<tr>
<td>Олег Піцун</td>
<td>West Ukrainian National University</td>
<td><a href="mailto:p.pitsun@wunu.edu.ua">p.pitsun@wunu.edu.ua</a></td>
<td>57190575875</td>
<td>T-3537-2019</td>
<td><a href="https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&amp;hl=en">https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>Тамара Дацко</td>
<td>I. Horbachevsky Ternopil National Medical University</td>
<td><a href="mailto:datsko_t@tdmu.edu.ua">datsko_t@tdmu.edu.ua</a></td>
<td>57188574014</td>
<td>57190575875</td>
<td><a href="https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&amp;hl=en">https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>Богдан Дериш</td>
<td>West Ukrainian National University</td>
<td><a href="mailto:d.dbb@wunu.edu.ua">d.dbb@wunu.edu.ua</a></td>
<td>57219592005</td>
<td>AAX-2580-2020</td>
<td><a href="https://scholar.google.com.ua/citations?hl=uk&amp;user=JJuCMAAAAJ">https://scholar.google.com.ua/citations?hl=uk&amp;user=JJuCMAAAAJ</a></td>
</tr>
<tr>
<td>Григорій Мельник</td>
<td>West Ukrainian National University</td>
<td><a href="mailto:mgm@wunu.edu.ua">mgm@wunu.edu.ua</a></td>
<td>27867794600</td>
<td>57219592005</td>
<td><a href="https://scholar.google.com.ua/citations?user=GmIyuE4AAAAJ&amp;hl=en">https://scholar.google.com.ua/citations?user=GmIyuE4AAAAJ&amp;hl=en</a></td>
</tr>
</tbody>
</table>

Олег Піцун, Олєг Піцун
West Ukrainian National University
e-mail: p.pitsun@wunu.edu.ua
https://orcid.org/0000-0003-6920-8786
Scopus Author ID: 57190575875
ResearcherID: T-3537-2019
https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&hl=en

Тамара Дацко
Tamara Datsko
I. Horbachevsky Ternopil National Medical University
e-mail: datsko_t@tdmu.edu.ua
https://orcid.org/0000-0001-9383-2629
Scopus Author ID: 57188574014
https://scholar.google.com.ua/citations?user=sP0KhEAAAAJ&hl=en

Богдан Дериш
Bohdan Derysh
West Ukrainian National University
e-mail: d.dbb@wunu.edu.ua
https://orcid.org/0000-0002-7215-9032
Scopus Author ID: 57219592005
ResearcherID: AAX-2580-2020
https://scholar.google.com.ua/citations?hl=uk&user=JJuCMAAAAJ

Григорій Мельник
Grygory Melnyk
West Ukrainian National University
e-mail: mgm@wunu.edu.ua
https://orcid.org/0000-0003-0646-7448
Scopus Author ID: 27867794600
ResearcherID: c-4526-2016
https://scholar.google.com.ua/citations?user=GmIyuE4AAAAJ&hl=en

кандидат технічних наук, старший викладач кафедри комп’ютерної інженерії Західноукраїнського національного університету. м. Тернопіль. Україна.

кандидат медичних наук, доцент кафедри патологічної анатомії Тернопільського національного медичного університету ім І. Я. Горбачевського. м. Тернопіль. Україна.

аспірант, викладач кафедри комп’ютерної інженерії Західноукраїнського національного університету. м. Тернопіль. Україна.

кандидат технічних наук, доцент кафедри комп’ютерної інженерії Західноукраїнського національного університету. м. Тернопіль. Україна.