Biomedical Engineering and Technology
Issue 11(3), 2023 ISSN (online) 2707-8434

UDC 004.81 + 616-006

КЛАСИФІКАЦІЯ ТУБЕРКУЛЬОЗНИХ УРАЖЕНЬ ЛЕГЕНЬ МЕТОДОМ ПОЗИЦІЙНОГО ГОЛОСУВАННЯ ЗА ДАНИМИ КОМП'ЮТЕРНОЇ ТОМОГРАФІЇ

Матвійчук Олександр
matviichuk.oleksandr@lll.kpi.ua
Настенко Євген Арнольдович
nastenko.e@gmail.com
кафедра біомедичної кібернетики
Національний технічний університет України
«;Київський політехнічний інститут імені Ігоря Сікорського»,
м. Київ, Україна

Анотація *– У дослідженні розглядається розробка процесу класифікації хіміочутливого та хіміорезистентного туберкульозу. Система що реалізує даний процес складається з двох етапів: відбору інформативного ансамблю ознак та навчання класифікатора. Відбір інформативного ансамблю ознак відбувається на зображеннях комп'ютерної томографії легень за допомогою матриць текстурних характеристик. Отримані ознаки фільтруються методом клас орієнтованої селекції в інформативний ансамбль. Навчання класифікатора "Random Forest" відбувається на сформованому селекцією ансамблі. До методу голосування "Random Forest" запропоновано покращення, яке оптимізує структуру та параметри функції голосування, та персоналізує сформований колектив голосуючих експертів .Дана система голосування збільшує точність класифікації на 5%, Система класифікації на виділених областях інтересу досягла точності у 88%. Результати демонструють ефективність реалізованого рішення при розв'язанні задачі класифікації типів ураження легень: «хіміочутливий», «хіміорезистентний»*

Ключові слова*: текстурний аналіз, виявлення легеневих патологій, томографія, фільтрація ознак, медичні зображення.*

I. ВСТУП

Туберкульоз легень – одне з найбільш поширених та небезпечних захворювань, що за даними Українського центру здоров'я зустрічається в 45 громадян на 100 000 [1]. Раннє та точне виявлення туберкульозних уражень є однією з умов успішного лікування даної хвороби. На сьогоднішній день застосовується декілька типів інструментів скринінгу для виявлення патологій легень [2]. Одним з інструментів є комп'ютерна томографія (КТ), результат якої доступний в цифровому вигляді та може бути оброблений методами машинного навчання.

Використання даних КТ для скринінгу туберкульозних захворювань розглядається в роботах [3, 4, 5, 6, 7]. Дослідження [3] за допомогою текстурного аналізу та клас-орієнтованої селекції вирішує задачу диференціації хіміочутливого та хіміорезистентного туберкульозу з точністю 83%. В роботі [4] аналізується зміна яскравості на границях легень з побудовою граничних карт. Подальший аналіз карт

дозволяє навчити класифікатор з точністю в 86.3% Використання комбінованих ознак отриманих статистичними та нейромережевими методами в [5] та нейронна мережа як класифікатор, дозволила отримати точність класифікації в 84.1%. В роботі [6] ознаки з матриць текстурних характеристик подаються нейронним мережам зі згортковою архітектурою та класифікаторами: support vector machine, k-nearest neighbors, random forest та multilayer perceptron. Найкращі результати класифікації на даних ознаках демонструють support vector machine (74.5%) та random forest (73.2%). Використання згорткових нейромереж без використання текстурного аналізу в [7] демонструє точність в 82.6%.

Дослідниками в [3, 5, 6] застосовано текстурний аналіз для формування набору вхідних ознак. Така попередня обробка зображень дозволяє отримати кращі результати класифікації в задачі хіміочутливий-хіміорезистентний туберкульоз.

Наведений аналіз робіт показує різноманіття методів для вирішення задачі класифікації хіміорезистентної форми туберкульозу легень, та актуальність пошуку засобів підвищення діагностичної точності КТ діагностики даного захворювання.

II. МЕТА ДОСЛІДЖЕННЯ

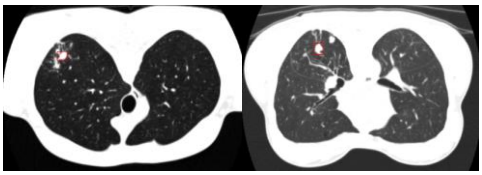
Підвищення точності класифікації в задачі диференціації легеневих форм туберкульозу по КТ зображенням.

III. МАТЕРІАЛИ ТА МЕТОДИ

Цей розділ описує етапи обробки в системі, починаючи з відбору ознак в «легеневому» вікні за допомогою методів віконної фільтрації та нормалізації зображень, вибору інформативного ансамблю ознак до класифікації, яка надає користувачеві клас до якого належить виділена область інтересу.

У досліджені використано анонімізовані дані КТ обстежень легень пацієнтів ДУ «Національний інститут фтизіатрії і пульмонології ім. Ф. Г. Яновського» НАМН України в рамках договору про співробітництво з НТУУ «КПІ ім. Ігоря Сікорського.

Дані з 15 серій знімків у форматі Neuroimaging Informatics Technology Initiative (NIfTI) надані та розмічені фахівцями. Знімки містять 853 області інтересу, з них 420 відносяться до хіміочутливого класу, а 433 до хіміорезистентного (рисунок 1). Для навчання, валідації та тестування відібрані набори розподілені у співвідношенні 70/20/10 (таблиця 1). Дані зображень комп'ютерної томографії представлено в одиницях густини матеріалів (одиниці Гаунсфілда).



А Б

Рисунок 1. Приклад виділеного регіону ураження легеневим туберкульозом. А – хіміочутливим, Б – хіміорезистентним.

Таблиця 1. Розбиття вхідного набору даних на області інтересу(шт)

Клас	*Навчання*	*Валідація*	*Тестування*	*Загалом*
Хіміорезистентний	*294*	*84*	*42*	*420*
Хіміочутливий	*304*	*86*	*43*	*433*

Методами машинного навчання дані зображень комп'ютерної томографії нормалізовано відносно «легеневого» вікна за областю дослідження, застосовано вікно з центром у «-600», та шириною у 1600 одиниць Гаунсфілда [7]. Над зображенням

вікна виконується операція нормалізації та конвертації яскравостей зображення в 256 градацій сірого. Конвертація зменшує розмірність даних для проведення текстурного аналізу, об'єднуючи значення сусідніх густин матеріалу.

Селекція текстурних ознак. За нормалізованими зображеннями кожної області інтересу побудовано матриці текстурних характеристик: gray level co-occurrence matrix (GLCM) [8], gray-level run length matrix (GLRLM) [9], gray level area size matrix (GLSZM) [10], neighboring gray level dependence matrix (NGTDM) [11], grey level difference matrix (GLDM) [12]. Кожна матриця текстурних характеристик формується незалежно. З метою оптимізації обробки зображення частина процесу - перетворення та селекція ознак виконується паралельно (канальна фільтрація). Паралелізації піддаються всі операції фільтрації в межах однієї матриці текстурного аналізу. До таких операцій відноситься перший етап клас-орієнтованої селекції, який складається з фільтрів за критерієм міжкласової, внутрішньокласової дисперсії та процедури крокового об'єднання. Фільтр за критерієм міжкласової дисперсії орієнтований на здатність розділити класи за кожною з ознак. Крокове об'єднання згортає сусідні ознаки та зменшує розмірність даних. Фільтр за максимізацією міжкласової та мінімізацією внутрішньокласової відстані орієнтований на стиснення класів та максимальну роздільність їх об'єктів. Ознаки, які не зустрілись в обох класах вилючаються на першому етапі виконання алгоритму, оскільки між ними не існує міжкласової відстані. Після канальної фільтрації відібрані з кожної матриці текстурних характеристик ознаки об'єднуються у єдиний ансамбль. Об'єднання виконується генетичним алгоритмом за критерієм кореляції з класами найбільш попарно незалежних ознак.

Класифікація ознак випадковим лісом. На об'єднаному ансамблі текстурних характеристик навчено декілька реалізацій алгоритму класифікації у класі методу «випадковий ліс». Випадковий ліс обрано, зважаючи на стабільність результатів, здатність до оптимізації структури дерев та високу точність в задачах класифікації [13]. Функція голосування класифікатору використовує метод колективного прийняття рішень, який базується на окремих класифікаторах - деревах. Дерева будуються в процесі бутстрепінгу та навчаються на випадкових піднаборах даних, що дозволяє боротись з перенавчанням моделі. Для класифікації нового об'єкта алгоритм застосовує раніше навчені дерева та повертає результат, який найчастіше зустрічається серед дерев.

На сформованому з областей інтересу ансамблі ознак бібліотечна версія випадкового лісу досягла загальної точності класифікації (overall accuracy) у 83,5%. Загальну точність класифікації в таблиці розраховано за формулою 1.

$$acc = \frac{1}{2}\left(\frac{N_1^+}{N_1} + \frac{N_2^+}{N_2}\right) \quad (1)$$

де, acc – точність класифікації, N_1^+, N_2^+ – кількість правильно класифікованих областей інтересу відповідного класу, N_1^+, N_2^+ – кількість областей інтересу в кожному класі.

Наступна версія класифікатору застосовує розширення простору ознак узагальненими змінними на вході дерев випадкового лісу, визначає структуру та значення вагових коефіцієнтів функції голосування дерев у лісі методом групового урахування аргументів (МГУА)[14]. Загальна точність розробленого класифікатора досягнута у 85%.

Класифікація деревами випадкового лісу із застосуванням методу позиційного голосування.

Стандартна версія голосування у класифікаторі «Випадковий Ліс» застосовує

механізм колективного прийняття рішень за стратегією рівнозваженого вибору [15]. Застосування у функції голосування вагових коефіцієнтів, що розраховані методом групового урахування аргументів збільшило точність класифікації та ініціювало ідею про формування суб'єктів функції голосування за компетентністним принципом та застосування методу позиційного голосування [16]. Формалізуємо поняття компетентності стану класифікатору.

Нехай маємо дерево рішень T із заданою структурою S, що зв'язує елементи (вузли) E дерева. Станом $St_j, j=1,n$ позначимо послідовність елементів $E_j=(E_0,E_{i_1},...,E_{i_{h_j}})_j$ задіяних при формуванні конкретного j-ого шляху при ухваленні рішення по дереву T. Тоді для кожного шляху $St_j, j=1,n$ можливо здійснити процедуру верифікації на верифікаційній вибірці об'єктів $V_i, i=1,v$. Прийнятий у конкретній задачі критерій якості рішення на даній верифікаційній вибірці назвемо компетенцією K_j стану (шляху по дереву) $St_j, j=1,n$. Тут E_0 – корінь дерева, $E_{i_{h_j}}$ - кінцевий елемент дерева, що завершує j-ий шлях по дереву прийняття рішень, h_j - кількість елементів (вузлів) дерева на j - тому шляху прийняття рішень, n - кількість шляхів прийняття рішень у дереві T, v - кількість об'єктів верифікаційної вибірки.

Формування оцінок за компетентністю виконується для заміщеної голосуючої сутності класифікатора з дерева на кінцеву вітку, що представляє собою маршрут по дереву. Кожну таку гілку (від кореня до кінцевого листка) ми назвали вище станом класифікатора, у якому приймається рішення для деякого підкласу об'єктів. З точки зору теорії прийняття рішень таку сутність називають експертом. Кожен експерт отримує характеристику «компетентність», що визначається відсотком співпадіння його голосу з правильним класом на валідаційній вибірці об'єктів. Визначення вище характеристики «компетентність» стану класифікатора (експерта) дозволяє сформувати функцію голосування за одним, найбільш компетентним експертом для кожного з об'єктів, що має бути класифікованим. Одержана загальна точність класифікації при такому підході становила 79.4%.

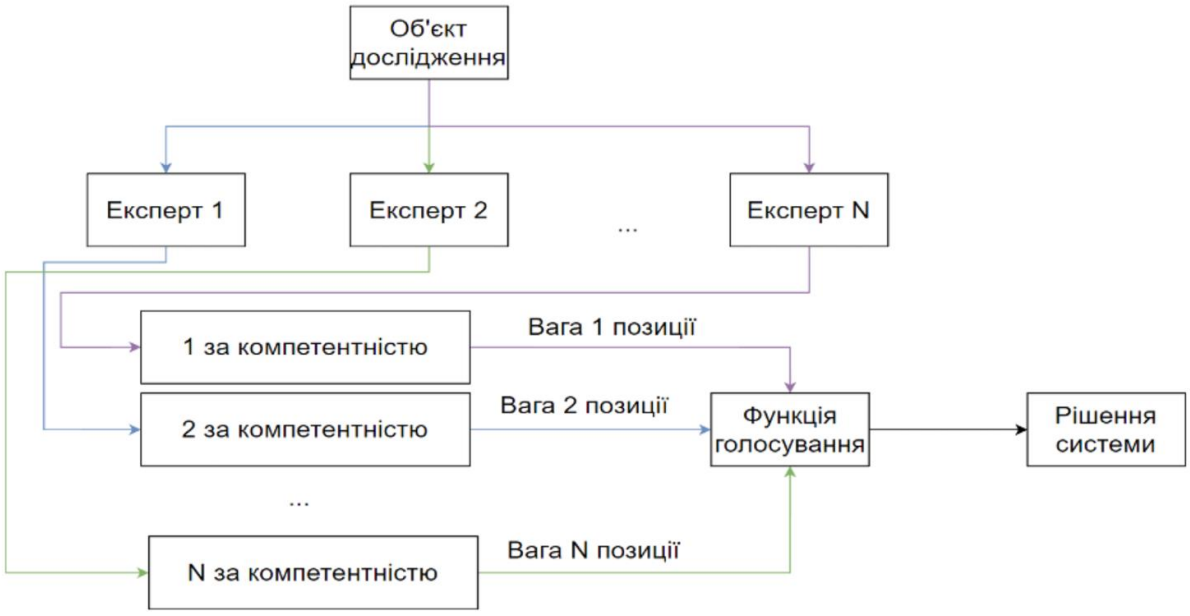


Рисунок 2. Схема голосування за позиціями експертів.

Наступним кроком до нової технології згортки голосуючих суб'єктів стала гіпотеза про вплив позицій компетентностей експертів та форми їх об'єднання на точність класифікації. Концепція пропонує формування віртуальних експертів, для яких визначається результат класифікації за фіксованим порядком компетентності. Статистика голосування віртуальних експертів, кожен з яких розміщується у порядку компетентності, передається в МГУА для визначення ваги такого «порядкового» експерта. В результаті, кожна позиція в голосуванні отримує свою вагу. Таким чином для одержання нової функції голосування спочатку формується віртуальна множина експертів, що приймають участь в голосуванні у порядку компетенції. Для кожного об'єкту навчальної та тестової вибірки обчислюється відповідна статистика для розрахунку за МГУА структури та параметрів функції голосування.

В режимі класифікації нового об'єкту формується відповідний власний набір експертів. Сформований набір впорядковується за компетентністю і відповідно до їх порядку застосовуються ваги позиції. Оскільки експерти обираються індивідуально для кожного вхідного об'єкта, результат голосування є персоналізованим, тобто найкращим, який можна отримати саме для цього об'єкта з голосування членів колективу вирішальних правил. Схема голосування за позиційним методом на прикладі дерев випадкового лісу представлена на рисунку 2. Загальна точність класифікації - 88,2%.

IV. РЕЗУЛЬТАТИ ТА ЇХ ОБГОВОРЕННЯ

В роботі представлено проблему машинного навчання пов'язану з класифікацією зображень, на прикладі вирішення задачі диференціації хіміочутливої та хіміорезистентної форми туберкульозу. Задачу вирішено використанням клас-орієнтованої технології селекції ознак отриманих з матриць текстурних характеристик та класифікатором "Random Forest" з удосконаленням функції голосування позиційним підходом.

Точність методів 1-4 розраховано на тестовій частині вхідної вибірки областей інтересу, наданих фахівцями ДУ "Національний інститут фтизіатрії і пульмонології ім. Ф. Г. Яновського НАМН України. Метод 1 використовує клас-орієнтовану селекцію для визначення

інформативного ансамблю текстурних ознак та стандартну версію випадкового лісу в якості класифікатора; метод 2 використовує визначений вище інформативний ансамбль текстурних ознак та випадковий ліс з оптимізацією функції голосування деревами за МГУА; метод 3 використовує визначений інформативний ансамбль текстурних ознак та рішення випадкового лісу за результатом класифікації самого компетентного експерту (вітки дерева) для відповідного об'єкту; метод 4 використовує визначений вище інформативний ансамбль текстурних ознак та рішення випадкового лісу з позиційним голосуванням за визначеною МГУА структурою та параметрами функції голосування; метод 5 демонструє результат, що використовує архітектуру згорткової нейронної мережі Unet [7].

Таблиця 2. Точність класифікації в задачі класифікації хіміочутливий-хіміорезистентний на тестовому наборі даних

Метод	Клас		Загальна
	Хіміочутливий	Хіміорезистентний	
1	84.2%	82.8%	83.5%
2	83.7%	86.3%	85%
3	80.5%	78.2%	79.4%
4	89%	87.4%	88.2%
5	85%	79%	82%

Клас-орієнтована селекція ознак відбирає та формує інформативний ансамбль, який з бібліотечною реалізацією класифікатора випадкового лісу демонструє точність в 83.5%. Отриманий результат демонструє інформативність сформованого ансамблю. Інформативність матриць текстурних характеристик проаналізовано в роботах [3, 5, 6], які демонструють схожі точності класифікації.

Використання МГУА для оптимізації функції голосування випадкового лісу збільшує точність класифікації на інформативному ансамблі.

Сформована гіпотеза про можливість покращити результат класифікації за допомогою удосконалення функції голосування випадкового лісу. Оптимізація функції голосування з виділенням з дерева найбільш компетентного експерта, та його навчання зменшує точність класифікації до 79.4%. Таке падіння точності демонструє важливість групового голосування.

Запропонований метод позиційного голосування. Для даного класу моделей введено поняття компетенції стану моделі, та розглянуто формалізацію задачі на підкласах – деревах прийняття рішень та нейронних мережах. Використання позиційного підходу збільшує точність голосування до 88.2%.

Отриманий результат демонструє ефективність технології в порівнянні з роботою [3] за рахунок покращення функції голосування. Робота [5], яка вирішує задачу класифікації за допомогою згорткової нейронної мережі демонструє меншу точність класифікації.

Робота [7], яка опирається на використання нейронної мережі з згортковою архітектурою формує власний ансамбль ознак в процесі свого навчання, який є менш інформативний.

V. ВИСНОВКИ

Розроблено автоматизовану систему диференціації хіміорезистентної форми туберкульозу за КТ-зображеннями легень. Система демонструє збільшення точності з 83.5% до 88.2% в порівнянні з статистичним методом [3]. Точність методу є вищою в порівнянні з роботами [3, 4, 5, 6, 7].

Удосконалення методу досягнуто використанням інформативного ансамблю текстурних ознак та класифікаторів - дерев випадкового лісу із застосуванням методу позиційного голосування з визначенням структури та параметрів функції голосування за МГУА. Інформаційний ансамбль ознак одержано клас-орієнтованою технологією селекції, що ґрунтується на попередньому паралельному відборі ознак за критеріями міжкласової та внутрішньокласової дисперсії та остаточною селекцією перспективних ансамблів через генетичний відбір за

комбінаційним критерієм: кореляція з класами найбільш попарно незалежних ознак.

Розроблена в процесі дослідження технологія може бути використана до інших завдань машинного навчання, в яких присутні фільтрація ознак та класифікація. Такий результат досягнуто завдяки використанню клас-орієнтованої селекції ознак та позиційного голосування колективом вирішальних правил. Отримані результати підтверджують наявність діагностичної інформації для побудови класифікаційних моделей на КТ зображеннях легень. Виявлення ознак хіміорезистентності туберкульозу на етапі КТ діагностики сприяє більш ранньому початку терапевтичних процедур і може позитивно впливати на тривалість терапії та тяжкість перебігу захворювання.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

ORCID ID та внесок авторів:
0000-0002-8988-099 О. В. Матвійчук(А)
0000-0002-1076-9337 Є. А. Настенко (Б)

А – Проектування програмного коду, паралелізація процесу аналізу текстурних ознак, розробка та застосування методів клас орієнтованої селекції, метод позиційного голосування, застосування МГУА до визначення структури та параметрів функції голосування.

Б – Концепція роботи, визначення компетенції стану моделі, оцінка методів фільтрації ознак, критичний огляд статті.

ПЕРЕЛІК ПОСИЛАНЬ

1. Tuberculosis statistic. Center for Public Health of the Ministry of Health of Ukraine. 2022; https://phc.org.ua/kontrol-zakhvoryuvan/tuberkuloz/statistika-z-tb.
2. Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches. Geneva: World Health Organization; 2016. 39 p.
3. Matviichuk O, Nosovets O, Linnik M, Davydko O, Pavlov P, Nastenko I. Class-Oriented Features Selection Technology in Medical Images Classification Problem on the Example of Distinguishing Between Tuberculosis Sensitive and Resistant Forms. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). 2021. pp. 385-389, doi: 10.1109/CSIT52700.2021.9648747.
4. Santosh K, Vajda S, Antani S, Thoma G. Edge map analysis in chest x-rays for automatic pulmonary abnormality screening. International Journal of Computer Assisted Radiology and Surgery, 11(9). 2016. pp. 1637–1646. doi:10.1007/s11548-016-1359-6.
5. Vajda S, Karargyris A, Jaeger S, Santosh K, Candemir S, Xue Z, Antani S, Thoma G. Feature selection for automatic tuberculosis screening in frontal chest radiographs. Journal of Medical Systems, 42(8). 2018. doi:10.1007/s10916-018-0991-9.
6. Karki M, Kantipudi K, Yang F, Yu H, Wang Y, Yaniv Z, Jaeger S. Generalization challenges in drug-resistant tuberculosis detection from chest x-rays. Diagnostics 2022, 12. 2022. pp 188-211. doi: 10.3390/diagnostics12010188.
7. Bondina MM, Kalmychkov AS, Kriventsov VE. Comparative analysis of algorithms filtration of medical images. Herald of the National Technical University KhPI Subject Issue Information Science and Modelling. 2012 ;38:14-25. DOI: 10.20535/ibb.2021.5.2.233051.
8. Hu, Zheng, “A GLCM Embedded CNN Strategy for Computer-aided Diagnosis in Intracerebral Hemorrhage”, arXiv arXiv:1906.02040, 2019.
9. Galloway, “Texture analysis using gray level run lengths’’, Computer Graphics and Image Processing, 4, pp.172-179, 1975. 10.1016/S0146-664X(75)80008-6.
10. Thibault, Fertil, Navarro, Pereira, Cau, Levy; Sequeira, Mari, “Texture Indexes and Gray Level Size Zone Matrix”, Application to Cell Nuclei Classification, Pattern Recognition and Information Processing (PRIP), pp. 140-145, 2009.
11. Sun, Wee, “Neighboring gray level dependence matrix for texture classification”, Computer Vision, Graphics, and Image Processing, 20, p.297, 1982. DOI: 10.1016/0146-664X(82)90093-4.
12. Rizal, Hidayat, Nugroho, “Modification of Grey Level Difference Matrix (GLDM) for Lung Sound Classification”, 2018 4th International Conference on Science and Technology (ICST), pp. 1-5, 2018. DOI: 10.1109/ICSTC.2018.8528650.
13. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). DOI: 10.1023/A:1010933404324.
14. Степашко В, Булгакова О, Зосімов В. Ітераційні алгоритми індуктивного моделювання. Наукова думка, 2018.ISBN: 978-966-00-1610-1
15. Holliday, Pacuit. Stable Voting. Constitutional Political Economy. 2023. DOI: 10.1007/s10602-022-09383-9.
16. Matviichuk, Biloshytska, Horodetska, Pavlov, Linnik, Nastenko. Positional Approach to the Voting Function Formation of Random Forest Trees as an Example of Solving the Differentiating Tuberculosis Forms Problem. 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT). DOI: 10.1109/CSIT56902.2022.10000450.

Biomedical Engineering and Technology
Issue 11(3), 2023 ISSN (online) 2707-8434

UDC 004.81 + 616-006

CLASSIFICATION OF TUBERCULOUS LUNG PATHOLOGY BY POSITIONAL VOTING ON COMPUTED TOMOGRAPHY IMAGES

Oleksandr Matviichuk
matviichuk.oleksandr@lll.kpi.ua
Ievgen Nastenko
nastenko.e@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Abstract - *This research deals with the development of a classification process for chemosensitive and chemoresistant tuberculosis. The system that implements this process consists of two stages: selection of an informative ensemble of features and classifier training. The informative feature set is selected from computed tomography images of the lungs using texture characteristic matrices. The obtained features are filtered by class-based selection into an informative ensemble. The Random Forest classifier is trained on the ensemble formed by selection. An improvement to the "Random Forest" voting method is proposed that optimizes the structure and parameters of the voting function and personalizes the formed team of voting experts. This voting system increases the classification accuracy by 5%, and the classification system on the selected areas of interest achieved an accuracy of 88%. The results demonstrate the effectiveness of the implemented solution in solving the problem of classifying types of lung lesions: "chemosensitive", "chemoresistant"*

Key words: *texture analysis, detection of pulmonary pathologies, tomography, feature filtering, medical images*