Journal of Natural Sciences and Engineering

Publisher

International Burch University, Sarajevo, Bosnia and Herzegovina

Editorial Office and Administration

International Burch University

Francuske revolucije bb, 71210 Ilidža, Sarajevo, Bosnia and Herzegovina

Website: https://www.ibu.edu.ba/journal-of-natural-sciences-and-engineering/

ISSN 2637-2835 (Print)

C O N T E N T

Respected readers,

Journal of Natural Sciences and Engineering (JONSAE) is a peer-reviewed biannual journal that aims at the publication and dissemination of original research articles on the latest developments in the fundamental theory, practice and application of engineering, science, and technology. We provide a platform for researchers, academicians, professionals, practitioners, and students to impart and share knowledge in the form of high-quality empirical and theoretical research papers. The journal covers all areas of genetics and bioengineering, electrical and electronics engineering, information technology, architecture, applied mathematics, computer sciences, and civil engineering.

In this second issue in 2022, I would like to express my gratitude to our authors and manuscript reviewers for their unselfish effort. You have continued your commitment and diligence to the Journal in helping us to produce quality and meaningful content that has the opportunity to advance the field. All of us had personal and professional challenges due to return to in-class teaching and dynamic experiences we have all witnessed in terms of changes in how the higher education functions in the post-pandemic period, and yet despite this, we managed to assure continuous publication of our Journal. Thank you all for being part of this wonderful academic endeavor.

This issue is comprised of three articles. A review article we are presenting you stands at the crossroad of biochemistry and molecular biology and deals with the current state of knowledge related to the hydrophobic interaction chromatography as an experimental method. Two original papers come from two different areas; one is dedicated to machine learning and another to solar irradiation.

In the end, we plan to build upon the excellent editorial infrastructure in the next years. In addition to managing the normal turnover of the Editorial Board members, we will seek to expand it by recruiting additional members who can provide expertise in areas that are not currently well represented. In particular, we hope to recruit board members with expertise in such areas as statistical and biostatistical methods, computer science, bioengineering, and biomedical engineering, among others. We also hope to leverage the expertise of the Editorial Board members to train the next generation of scholars and potential Editorial Board members by finding ways to pair student reviewers with senior reviewers for a peer-review mentorship as a part of building a better research environment for new scientists. A similar approach is planned for the administrative members of the Editorial Board, with the aim of improving the quality of the overall Journal design and acquiring a modern visual identity.

Having in mind all stated, I wholeheartedly invite you to read this issue and join our team.

Yours sincerely,

Adna Ašić, PhD
Editor in Chief

Hydrophobic Interaction Chromatography: A Key Method for Protein Separation

Nermana Kovačević, Mohammad Tariq AL-Boush
International Burch University
Sarajevo, Bosnia, and Herzegovina
nermana.kovacevic@stu.ibu.edu.ba
tariq.boushi@stu.ibu.edu.ba

Literature review

Abstract: *This review's main goal is to increase theoretical knowledge and comprehension of the techniques used to separate important enzymes using chromatography. Protein separation, purification, and analysis frequently involve the use of hydrophobic interaction chromatography (HIC), in which the molecules are separated based on differences in hydrophobicity. It is conducted using a weakly hydrophobic, nonpolar stationary phase to which the proteins bind in an aqueous high-salt solution as the mobile phase. The protein integrity is thereby conserved during the process. HIC is usually used with a gradient from high to low concentrations of salt as an eluent and can be applied in a wide range of biotechnological processes. In this study, after a brief overview of hydrophobic interaction chromatography, we explain the HIC procedure, protein retention mechanism, factors affecting HIC, and applications of HIC.*

Keywords: Hydrophobic interaction chromatography, protein purification.

1. Introduction

Based on their hydrophobicity, biomolecules are separated and purified using the hydrophobic interaction chromatography (HIC) technique [17]. The term "salting-out chromatography" was initially used to describe this method in 1949 by Shepard and Tiselius [21]. Shaltiel and Er-el [20] renamed the process "hydrophobic chromatography" in 1973. Because proteins can stay on weakly hydrophobic matrices in the presence of salt, Hjerten (1973) coined the term "hydrophobic interaction chromatography" [11]. The efficacy of this approach has been the subject of numerous studies. Some research has found that this method can be used to separate and purify proteins in their native state [18], isolate protein complexes [5], and study protein folding and unfolding [3]. As we can see, this method is most commonly used for protein purification. Except for protein purification, hydrophobic interaction chromatography can be used to separate and purify cells, viruses, nucleic acids, and carbohydrates [6].

2. Hydrophobic Interaction Chromatography Procedure

In the hydrophobic interaction chromatography, the mobile phase passes through a stationary phase that contains a hydrophobic ligand. This allows the hydrophobic species found in the mobile phase to bind to the immobilized hydrophobic ligands found in the stationary phase [6]. HIC's stationary phase is made up of a resin matrix containing hydrophobic ligands. Linear chain alkyl groups such as ether, butyl, hexyl, and octyl are the most common and frequently used ligands in HIC [19]. Elution can be carried out by lowering the ionic strength of the mobile phase linearly or gradually and modifying the pH and temperature of the elution buffer [16]. The isolation of a particular HIC-compatible protein typically requires experimental research to identify the best chromatographic medium or buffer conditions. Among the commercially available HIC media, there are differences in the chemical composition of the functional groups, their hydrophobicity, density, and the size of the inert matrix beads to which the functional groups are linked. Unique chromatographic elution patterns are produced by these changes in the HIC medium.

Functional group densities range from 5 to 50 μmol/ml of medium, with matrix bead sizes ranging from 30 to 100 μm. Relatively small bead diameters and higher functional group densities provide better chromatographic resolution, as opposed to larger bead diameters

and lower functional group densities. Those operating conditions are suggested for highly concentrated mixture separations and faster flow rates. Numerous different factors that are routinely improved to enhance HIC protein purification elution features include elution buffer molarity, pH, and chromatographic flow rate [22]. Furthermore, the hydrophobic nature of the resin, the nature and structure of the protein sample, the prevalence and distribution of surface-exposed hydrophobic residues, and the type and quantity of salt in an aqueous binding buffer also have an important role in the separation of biomolecules by HIC [8].

The resolution of hydrophobic interaction chromatography is usually satisfactory. To improve resolution, it is critical to experiment with different operating conditions. Jennissen (2000) [13], suggested using the crucial hydrophobicity approach, which consists of three main steps: choosing an acceptable alkyl chain length, determining the critical surface concentration of alkyl residues, and determining the minimal salt concentration needed to achieve complete adsorption of proteins [15].

3. Hydrophobic Interaction

In an aqueous environment, hydrophobicity is defined as the association of nonpolar molecules that results from water's tendency to reject nonpolar molecules. Two hydrophobic or non-polar molecules will cluster when discovered in a polar environment to reduce contact with the polar solvent, such as water, methanol, etc. "Hydrophobic interaction" is the term given to this process. In biological systems, those interactions are extremely important. Hydrophobic interactions are important in many biological procedures, such as in antibody-antigen interactions, enzyme catalysis, protein aggregation and regulation. Furthermore, they are one of the primary driving forces responsible for the stability and folding of protein structures [23].

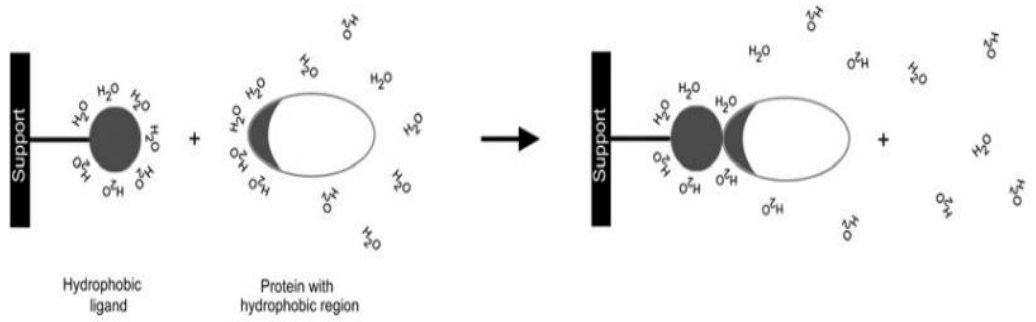

Figure 1. Illustration of hydrophobic interaction between immobilizing ligand and a protein [23]

4. Retention Mechanism in the HIC

In HIC, there is one very important mechanism known as a protein retention mechanism. This macromolecule retention occurs because of hydrophobic interaction. This interaction occurred between hydrophobic ligands which are found on a stationary phase and non-polar parts of the proteins, Figure 2. In HIC, several stationary phases are utilized, including agarose, polyacrylamide, cellulose, and dextran. Their primary characteristics include their high moisture absorption, high porosity, and potential for chemical modification.

Alkyl groups or aryl groups with 4–10 carbons act as ligands (weakly non-polar or hydrophobic) attached to the stationary phase. To prevent self-folding, there are no more than 10 carbon atoms in total [1]. In HIC, the hydrophobic ligands butyl (four carbons), octyl (eight carbons), and phenyl (the aromatic ring that encourages - interactions with the aromatic residues on the surface of a protein) are the most frequently utilized. Butyl is the least hydrophobic of the carbon chains used as HIC ligands because it is the shortest, octyl has an intermediate level of hydrophobicity, and phenyl displays the most hydrophobic interaction [1].

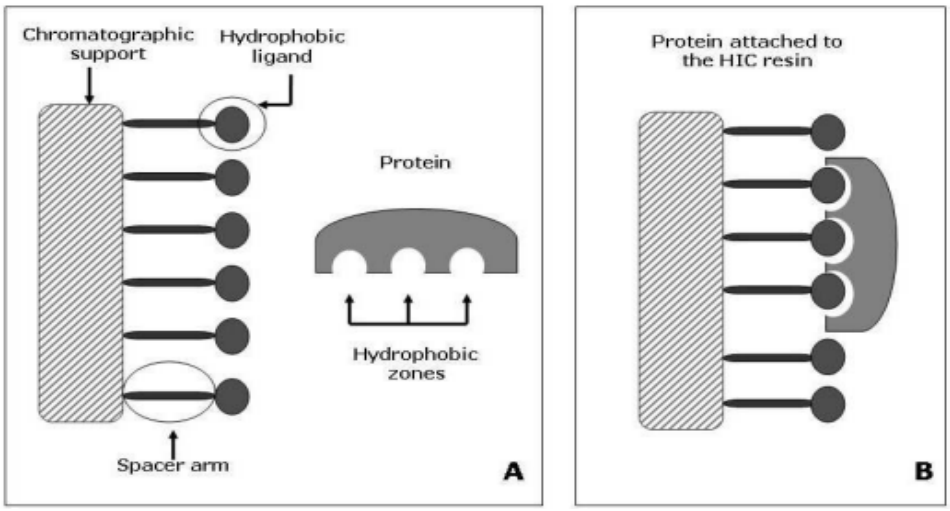

Figure 2. Retention mechanism in hydrophobic interaction chromatography [1]

The presence of neutral salts in the HIC enhances the protein retention mechanism. Whenever a neutral salt is added to a solution that contains a polar solvent, competition for the water molecules that hydrate the proteins occurs. This competition is more beneficial to salt. As a result, high salt concentration reduces the number of solvent molecules surrounding the protein, favoring the hydrophobic interaction between them. When that solution comes into contact with HIC resin, it enhances the interaction between

the proteins and the non-polar ligands on the resin surface. As a result, protein adsorption to the HIC stationary phase occurs. It is critical to select the appropriate salt type and concentration to reduce protein precipitation caused by solubility decrease in the presence of high salt concentration [1].

5. Factors that Impact HIC

A variety of factors influence HIC, including protein quality, matrix composition, ligand type and density, salt type and concentration, temperature, and pH. However, some factors, such as pH and temperature, are unpredictable and should be considered during technique optimization to improve selectivity, resolution, and binding capacity [8].

Type of ligands

The type of immobilizing ligands has a significant impact on the HIC adsorbent's selectivity for the protein. In hydrophobic interaction chromatography, alkyl and aryl chains that are covalently bonded to a base matrix are the most typical form of ligands [9]. While aryl ligands exhibit mixed mode behavior, which combines both aromatic and hydrophobic characteristics, alkyl ligands exhibit pure hydrophobic character [23]. The protein binding capacities of HIC adsorbents increase with longer alkyl chains at a constant degree of substitution [7]. For each specific circumstance, screening experiments should be used to determine whether to use aryl or alkyl ligands [23].

Base matrix

The matrix's characteristics are determined by its chemical structure and particle size. Different types of matrix materials can be used in HIC, like natural polymers such as cellulose, agarose, dextran, or chitosan, as well as synthetic polymers like polymethacrylate and inorganic compounds (silica). To enable reversible adsorption and prevent the HIC adsorbent from becoming much more hydrophobic, the matrix material must be hydrophilic. For reversible adsorption to be possible and for the matrix material not to considerably increase the HIC adsorbent's hydrophobicity, it must be hydrophilic. The most frequently used matrices in HIC are agarose or cellulose [23].

Type and Concentration of Salt

The Hofmeister series describes how ions affect hydrophobic interaction [12]. The salts at the beginning of the Hofmeister series, referred to as lyotropic salts, tend to induce

hydrophobic interactions by increasing "salting out" effects. Because of a "salting-in effect," the salts near the end (right) tend to reduce interactions. On the left side of Figure 3. anions and cations have a positive effect on HIC retention since they enhance ligand-protein interaction. On the other side anions and cations on the right side of Figure 3. despite increasing the surface tension of the water, do not promote HIC retention [23].

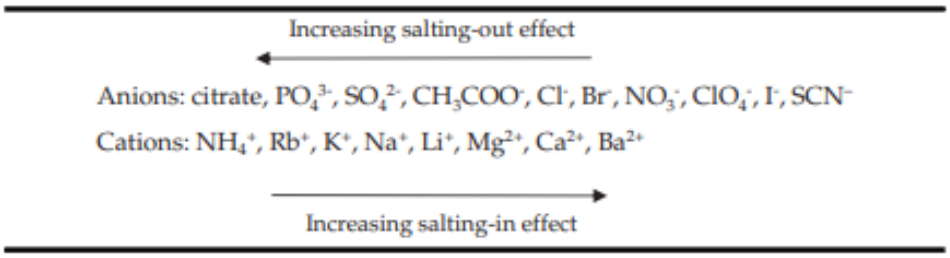

Figure 3. Hofmeister series. The impact of anions and cations on protein precipitation [20]

When a high concentration of lyotropic salts is added to a protein solution, the water molecules of the hydration shell may prefer to interact with the salt ions instead of the charged amino acid of the proteins. It thus exposes hydrophobic patches, which enable interaction between hydrophobic proteins or interaction between proteins ligands via Van der Waals interactions [10]. Among useful salts, sodium sulfate is the least viscous while sodium sulfate and sodium phosphate produce high molal surface tension. Ammonium sulfate is soluble in water and has high purity, but it can be degraded at an alkaline pH. Different salts can differently effect on retention of hydrophilic and hydrophobic proteins. In comparison to ammonium sulfate, sodium acetate tends to increase the retention mechanism of hydrophobic and hydrophilic proteins, whereas sodium citrate has the opposite effect [7].

pH

Higher pH tends to weaken hydrophobic interactions due to increased titration of charged groups and an increase in protein hydrophilicity. As a result, lower pH promotes hydrophobic interaction. Proteins that are unable to bind to HIC absorbents at neutral pH bind to them at acidic pH. Protein retention varies more noticeably at pH levels above 8.5 and below 5 than it does in the pH 5–8.5 range [11]. These results demonstrate that pH is a crucial separation factor for hydrophobic interaction chromatography optimization [9].

Temperature

Temperature increases usually have a positive effect on retention, while decreasing temperature favors elution. Also, protein conformational states and solubility may be affected by high temperatures. As a result, the temperature is frequently one of the characteristics that is kept relatively constant. However, in large-scale production, maintaining a constant temperature may be an expensive and difficult task. Retention in HIC was also discovered to be an entropy-driven process at low temperatures and an enthalpy-driven process at high temperatures [23].

Additives

Other methods can be applied in HIC for protein purification instead of lowering salt concentration [23]. In hydrophobic interaction chromatography, small concentrations of detergents, alcohols, and soluble compounds that possess salt promote a weakening of the interaction between proteins and ligands. This causes the bound solute to be deactivated and removed, allowing the elution of the desired product. The hydrophobic regions of alcohols and detergents compete with proteins for adsorption sites on the chromatography media. When highly hydrophobic proteins are linked to the gel medium, additives are also very important and efficient in cleaning HIC columns [7].

6. Literature Review

Hydrophobic interaction chromatography has proven to be a successful method for the purification of enzymes, especially β-glucosidases from plants, animals, insects and fungi, but also many other enzymes**.**
Below are several studies showing the successfulness of this method as a means of protein analysis.

Koffi et al. (2012) article was a study aiming to purify β-glucosidase from cockroaches, Periplaneta Americana, using hydrophobic interaction chromatography and other methods. Periplaneta americana were captured from rooms. They were collected directly from the nest and then stored. The enzyme was purified 9.93-fold to a specific activity of 40.33 (U/mg of protein) and an overall yield of 1.54 %. The optimal pH for studying Periplaneta americana β-glucosidase was 3.6, and the optimal temperature was 55 °C. The molecular weight for Periplaneta americana β-glucosidase was 43.8 kDa, meaning that the enzyme was monomeric (Table 1). The kinetic parameters km and Vma of β-glucosidase

were measured using three substrates: cellobiose, p-nitrophenyl-β-D-glucopyranoside, and pNP-N-acetyl-β-D-glucopyranoside. The km values for those substrates were found to be 3.29mM and 41.67 U/mg for pNP-β-D-Glucopyranoside, 0.37mM and 11.49 U/mg for pNP-N-acetyl-β-D-glucopyranoside and 8.52mM and 39.12 U/mg for cellobiose (Table 2). These findings reveal that the catalytic activity of β-glucosidase is significantly higher for pNP-N-acetyl—D-glucopyranoside than for cellobiose and pNP-D-glucopyranoside. Ultimately, as a result and conclusion, the article reported the successful isolation and purification of the enzyme using hydrophobic interaction chromatography [14].

Bešić et al. (2016) study aimed to isolate and purify β-glucosidase from brassica oleracea by salting out with ammonium sulfate and hydrophobic interaction chromatography. The main active fraction of the β-glucosidase was purified sevenfold with a yield of 4.1%. This study showed that β-glucosidase isolated from broccoli was a dimer (130 kD) made up of one major and one minor subunit (80 kD and 50 kD). Enzyme properties, such as the effect of different inhibitors, kinetic parameters, and optimum environmental parameters, were determined. The pH optimum was 6.0 and the temperature optimum was 35 °C (Table 1). The kinetic parameters km and Vmax of broccoli β-glucosidase were determined by using four substrates: 4-Nitrophenyl-b-D-glucopyranoside (p-NPG), ortho-Nitrophenyl-b-D-glucopyranoside (o-NPG), paraNitrophenyl-b-D-galactoside (p-NPGal), and ortho-Nitrophenyl-b-D-galact The km values for those four substrates (p-NPG, o-NPG, p-NPGal, and o-NPGal) were determined to be 0.755 mM, 0.174 mM, 0.988 mM, and 0.213 Mm, and the Vmax values were 604 U/mg, 38 U/mg, 556 U/mg, and 308 U/mg (Table 2). Considering that the Vmax values in all four instances were high, it is possible to conclude that β-glucosidase from broccoli has a strong affinity and interaction with those four substrates. Regarding inhibition experiments, p-NPG was used as a substrate and glucose and β–gluconolactone as inhibitors, with ki values of 0.64 mM and 0.038 Mm. A study showed that glucose and β–gluconolactone completely inhibit the broccoli β-glucosidase with ki values of 0.038 mM and 0.64 Mm, and inhibitions were competitive for both inhibitors. So this study eventually reported a successful purification of the beta-glucosidase using the hydrophobic interaction chromatography technique [4].

Table 1. Properties of β-glucosidase isolated using HIC

	Molecular mass (kDa)	Subunit molecular mass (kDa)	Quaternary structure	Optimum pH	Optimum temperature °C
Brassica oleracea	130	80 50	Dimer	6.0	35
Agaricus bisporus	110	46 62	Dimer	4.0	55
Periplaneta americana	43.8	-	Monomer	3.6	55

(-) Not determined

Ašić et al. (2015) study aimed to isolate and purify β-glucosidase from Agaricus bisporus (White Button Mushroom) using ammonium sulfate precipitation and hydrophobic interaction chromatography. Agaricus bisporus β-glucosidase was purified 10.12-fold during the precipitation and chromatography steps. They found that the enzyme was a dimer with two subunits of approximately 46 and 62 kDa. The enzyme functions best at a pH of 4.0 and a temperature of 55°C (Table 1). The enzyme was found to be exceptionally thermostable. To study enzyme activity, two substrates were used: p-NPGlu and o-NPGlu. The km values for those substrates were found to be 1.751 mM and 8.547 mM, and the Vmax values were 833 U/mg and 556 U/mg (Table 2). As compared to o-NPGlu, A. bisporus exhibits a much stronger affinity for p-NPGlu as a substrate. A. bisporus β-glucosidase was inhibited by both gluconolactone and glucose. Both function as competitive inhibitors, with gluconolactone being a considerably more effective inhibitor. This is supported by a comparison of the computed Ki values for glucose and gluconolactone, which were 9.402 mM for glucose and 0.0072 mM for gluconolactone. The β-glucosidase from Agaricus bisporus was successfully purified and biochemically characterized using ammonium sulfate precipitation and hydrophobic interaction chromatography [2].

Table 2. Kinetic parameters of β-glucosidases

	Substrate	Km (mM)	Vmax
Brassica oleracea	p-NPG	0.755	604 U/mg
	o-NPG	0.174	38 U/mg
	p-NPGal	0.988	556 U/mg
	o-NPGal	0.213	308 U/mg
	o-NPG	14.11	48.5 U/mg
Agaricus bisporus	p-NPGlu	1.751	833 U/mg
	o-NPGlu	8.547	556 U/mg
Periplaneta americana	pNP-beta-D-Glucopyranoside	3.29	41.67 Umg
	pNP-N-acetyl-beta-D-Glucopyranoside	0.37	11.49 U/mg
	Cellobiose	8.52	39.12 U/mg

7. Applications of HIC

HIC is frequently used in the production of highly purified biomedical products such as therapeutic proteins, monoclonal antibodies, and enzymes. To purify target proteins, hydrophobic interaction chromatography can be used as a single step or together with other chromatography methods. HIC is also a particularly helpful method in large-scale industrial applications. to purify antibodies.

Purification of plasmid by hydrophobic interaction chromatography was achieved using sodium citrate-based buffers. Successful HIC of ribonuclease A, ovalbumin, and lactoglobulin was performed at alkaline pH (9.5) using monosodium glutamate. Purification of monoclonal antibodies was successfully achieved by HIC. HIC was used to separate lysozyme from chicken egg white. Hydrophobic interaction chromatography was used to purify human PON1Q192 and PON1R192 isoenzymes. HIC together with ion exchange chromatography and ammonium sulfate are used for successful purification of recombinant HIV reverse transcriptase [5].

8. Conclusion

In conclusion, in biological systems, hydrophobic interactions are extremely important. They are the most critical factor in protein folding and structural stabilization, as well as other biological procedures such as in reactions between antibodies and antigens. HIC uses the protein's hydrophobicity to promote separation via hydrophobic interactions between non-polar ligands and hydrophobic areas on the proteins [21]. Hydrophobic interaction chromatography is currently a widely-used and effective separation method for enzyme purification on a laboratory and industrial scale.

References

1. Andrea Mahn (2012). Hydrophobic Interaction Chromatography: Fundamentals and Applications in Biomedical Engineering, Biomedical Science, Engineering and Technology, Prof. Dhanjoo N. Ghista (Ed.), ISBN: 978-953- 307-471-9, InTech.
2. Ašić, A., Bešić, L., Muhović, I., Dogan, S., & Turan, Y. (2015). Purification and characterization of β-glucosidase from Agaricus bisporus (white button mushroom). The Protein Journal, 34(6), 453-461.
3. Bai, Q., Wei, Y.M., Geng, M.H., Geng X.D. (1997). High performance hydrophobic interaction chromatography - A new approach to separate intermediates of protein

folding .1. Separation of intermediates of urea-unfolded alpha-amylase. Chinese Chemical Letters, 8, 67-70.

4. Bešić, L., Ašić, A., Muhović, I., Dogan, S., & Turan, Y. (2017). Purification and Characterization of β-Glucosidase from Brassica oleracea. Journal of Food Processing and Preservation, 41(2), e12764.
5. Chaturvedi, R., Bhakuni, V., Tuli, R. (2000). The delta-endotoxin proteins accumulate in Escherichia coli as a protein-DNA complex that can be dissociated by hydrophobic interaction chromatography. Protein Expression and Purification, 20, 21-26.
6. Charcosset, C. (2012). Membrane chromatography. Membrane Processes in Biotechnology and Pharmaceutics, 169–212. doi:10.1016/b978-0-444-56334-7.00005-8.
7. Desai Sonal. (2009). Hydrophobic Interaction Chromatography- An Important Technique for Separation of Proteins. International Journal of Pharmaceutical Research. 1. 40-49.
8. Eriksson, K. O. (2018). Hydrophobic Interaction Chromatography. Biopharmaceutical Processing, 401–408. doi:10.1016/b978-0-08-100623-8.00019-0.
9. Eriksson KO, Belew M.(2011). Hydrophobic interaction chromatography. Methods Biochem Anal.;54:165-81. doi: 10.1002/9780470939932.ch6. PMID: 21954777.
10. Fleming R. (2020). ADC Analysis by Hydrophobic Interaction Chromatography. Tumey LN, editor, Antibody-Drug Conjugates: Methods and Protocols, Springer US, New York, NY, Methods in Molecular Biology, 147–161.
11. Hjerten, S. (1973). Some general aspects of hydrophobic interaction chromatography. Journal of Chromatography, 87, 325-331.
12. Hofmeister, F. (1988). On regularities in the albumin precipitation reactions with salts and their relationship to physiological behavior. Arch. Exp. Pathol. Pharmakol. 24, 247–260.
13. Jennissen HP (2000) Hydrophobic interaction chromatography: the critical hydrophobicity approach. International Journal of BioChromatography 5: 131–163.
14. Koffi Y. G., Konan K. H., Kouadio E. J. P., Dabonn S., and Kouamé L. P. (2012). Purification and biochemical characterization of β-glucosidase from cockroach Periplaneta Americana. Journal of Animal and Plant Sciences, 13, 1747-1757.
15. Lienqueo, M. E., Mahn, A., Vásquez, L., & Asenjo, J. A. (2003). Methodology for predicting the separation of proteins by hydrophobic interaction chromatography and its application to a cell extract. Journal of Chromatography A, 1009(1-2), 189–196. doi:10.1016/s0021-9673(03)00924-5.

16. Lienqueo, M. E., Shene, C., & Asenjo, J. (2009). Optimization of hydrophobic interaction chromatography using a mathematical model of elution curves of a protein mixture. Journal of Molecular Recognition, 22(2), 110–120. doi:10.1002/jmr.927.
17. Liu, C.-I., Hsu, K.-Y., & Ruaan, R.-C. (2006). Hydrophobic Contribution of Amino Acids in Peptides Measured by Hydrophobic Interaction Chromatography. The Journal of Physical Chemistry B, 110(18), 91 48–9154. doi:10.1021/jp055382f.
18. Porath J, Sundberg L, Fornstedt N and Olsson I (1973) Salting-out in amphiphilic gels as a new approach to hydrophobic adsorption. Nature 245: 465–466.
19. Queiroz JA, Tomaz CT, Cabral JMS. (2001). Hydrophobic interaction chromatography of proteins. Journal of Biotechnology 87: 143–159.
20. Shaltiel, S., Er-el, Z. (1973). Hydrophobic chromatography: use for purification of glycogen synthetase. Proceedings of the National Academy of Sciences U.S.A., 70, 778- 781.
21. Shepard, C.C., Tiselius, A. (1949). In "Chromatographic Analysis" p. 275. Discussions of the Faraday Society, 7. Hazell, Watson and Winey. London.
22. Stone OJ, Biette KM, Murphy PJ. (2014). Semi-automated hydrophobic interaction chromatography column scouting used in the two-step purification of recombinant green fluorescent protein. PLoS One ;9(9):e108611. doi: 10.1371/journal.pone.0108611. PMID: 25254496; PMCID: PMC4177899.
23. Tomaz, C.T. & Queiroz J. A. (2013). In Liquid Chromatography: Fundamentals and Instrumentation (Eds.), Elsevier Inc.2013 pp. 121– 142.
24. Ueberbacher, R., Haimer, E., Hahn, R., & Jungbauer, A. (2008). Hydrophobic interaction chromatography of proteins: V. Quantitative assessment of conformational changes. Journal of Chromatography A, 1198, 154-163

Solar Irradiation Prediction Based on M5 Model Tree and Feature Importance Evaluation

Lejla Idrizović, Lejla Lulić Skopljak, Faris Haznadarević, Haris Ahmetović
International Burch University
Sarajevo, Bosnia and Herzegovina
lejla.idrizovic@stu.ibu.edu.ba
lejla.lulic.skopljak@stu.ibu.edu.ba
faris.haznadarevic@stu.ibu.edu.ba
haris.ahmetovic@ibu.edu.ba

Orginal research

Abstract: *In the last decade, the usage of renewable energy is on the rise, and that trend will only continue because technology is becoming more developed, so renewable energy sources are going to offer more for the same price. Besides all positive properties, there are also some negatives like direct dependence on the weather conditions. That means energy production is constantly changing, so it must be as precisely as possible predicted to be usable on a large scale. Fifteen attributes were analyzed using M5 regression tree. High positive degree of correlation was found between participle water and dew point temperature, air temperature with dew point, air temperature with precipitation of water, snow depth with Albedo daily, zenith angle with relative humidity, GHI with Air temperature. It was found that the zenith angle, between the normal of the Earth's surface and the Sun, was the most important feature of the dataset for solar irradiation prediction.*

Keywords: machine learning, photovoltaic, renewable energy, solar irradiation, weather forecast.

1. Introduction

Today, energy is increasingly used in all its forms, but conventional methods have had a severe effect on our environment. Therefore, new methods were developed. People are becoming more aware of environmental problems caused by the usage of fossil fuels. Therefore, there is a high demand for the usage of renewable energy. Limited supply, increasing costs, climate change concerns and government mandates are also driving a desire to increase the percentage of electricity generated by renewable energy sources. Besides all positive properties, there are also some negatives like direct dependence on the weather conditions. That means energy production is constantly changing, so it must be as precisely as possible predicted to be usable on a large scale. Currently, the largest and most widespread energy source used by humans in the world is fossil fuel. Their uncontrolled exploitation and use in the last century have caused extensive environmental pollution, caused by enormous production of greenhouse gases during exploitation and use, resulting in climate changes. Another reason for reducing reliance on fossil fuels is that they are limited, so they can be depleted very easily if they continue to be used to the present extent.

On the other hand, clean energy from the sun and wind is present everywhere around the world in unlimited quantities, and we only need to harvest and use that energy. The solution to the problem of environmental pollution and climate change is not and cannot be instantaneous, so the results will be visible over a longer period. Renewable energy sources are a feasible and cost-effective solution that when used for the generation of electric energy also introduces certain additional complications into existing electrical energy systems. To meet the needs of modern society for energy and to achieve sustainability it is necessary to make a planned transition from the use of fossil fuels to renewable energy sources [1]. Renewable electrical energy sources introduce additional complexity to the process of maintaining the power quality, stability, and reliability of electric power systems. The reason for that additional complexity is that renewable energy comes from natural sources (processes), which are intermittent by nature as shown in Figure 1, where we presented examples of Global Horizontal Irradiation (GHI) patterns through one day. Regarding that, we can conclude that renewable energy depends on weather conditions, and there are also important tasks of reducing that unpredictability [2]. That challenge can be solved or at least significantly reduced with the use of modern analyzing techniques of current and historic weather data with the goal of precise prediction of the weather conditions. Furthermore, the development of computer hardware with higher computational power enabled the use of large amounts of data for providing highly useful results using machine learning techniques.

The purpose of this paper is to give a contribution to further development and research of solar irradiation prediction using machine learning techniques. The need for the development of this topic is on the rise because renewable energy is everywhere around the world, and humans only need to harvest it in the right way and use it. If the use of renewable energy continues to grow, it will modernize the world's electricity grid, making it smarter, more secure, and better integrated around the world, also most important is that our environment and our lives will be much healthier and more sustainable.

2. Literature Review

In [3], authors applied multivariate adaptive regression tree, M5, and random forest models for solar irradiation prediction for 1-day to 6- day ahead hourly prediction. Nine variables, minimum temperature, maximum temperature, wind speed, rainfall, dew point, global solar irradiation, atmospheric pressure, and solar azimuth were used as the inputs for model creation. Authors used root means square error to determine which models provided the best results and the result was validated using the t-static error.

Authors in [4] employed different machine learning algorithms for the precise estimation of solar irradiation for two locations. Ten attributes, namely year, month, day, hour, pressure, temperature, humidity, wind speed, hourly solar duration, and solar irradiation were used for the best attribute selection using six different feature selection methods to create data for five selection groups. It was shown that hourly solar duration was the most important feature for both selected data groups.

In [5], authors used maximum temperature, minimum temperature, sunshine hours, wind speed and relative humidity for inputs to build models for estimating solar radiation. Kriging, response surface method, multivariate adaptive regression and M5 model tree were applied.

Meenal and Slevakumar [6] evaluated artificial neural networks (ANN), Support Vector Machine (SVM) and empirical solar radiation models with different combination models. Eight attributes, month, latitude, longitude, bright sunshine hours, day length, relative humidity, maximum and minimum temperature were used as parameters. Best attributes were determined, and the most important feature to reduce the dimensionality of the data was identified to improve the correlation coefficient and the prediction accuracy of solar irradiation models.

3. Data Collection

Data used in this research are from Solcast website, which provides solar forecasting data for free in limited amounts if they are not going to be used commercially. Solcast is providing historical forecasting weather data from the beginning of 2007. year to present days, extracted from the high-resolution satellite imagery using advanced modeling techniques. These images are taken using geostationary meteorological satellites of the newer generation capable of providing high-resolution (1-2km) imagery every 5 to 15 minutes. There are 5 different measurement resolutions from every 60 minutes to every 5 minutes, and all this data are in the Comma Separated Value (CSV) format [7].

Historical weather data for the location of one potential photovoltaic (PV) power plant close to the center of the capital city of Bosnia and Herzegovina, Sarajevo, (longitude of 18,444170° and latitude of 43,891336°) are used. The time period of the used data set is chosen to be the longest possible, which is the period from the beginning of the year 2007. until the time when this data set is downloaded during December of 2020. Within that wide time period, data set resolution is limited to 10 minutes, because of computational limits of hardware used for model calculation and because this data set already gave highly satisfactory results.

In the research, the prediction of Solar irradiation has been performed for the given attributes. The sun radiates over the given area and provides insight into the possible energy production of a PV system. Furthermore, with that information, the number of solar panels and their installed power can be estimated. The solar irradiation value that has been predicted in this study is more precisely the Global Horizontal Irradiation (GHI). GHI is the amount of direct and diffuse solar irradiance received on the horizontal surface, measured with the unit of watts per meter squared (W/m^2) [7]. In Figure 1, a few examples of a daily GHI's from one summer month with different weather conditions are shown. Case from an entirely clear sky sunny day (yellow) to partly sunny day (green), and a completely cloudy day are shown (gray and orange).

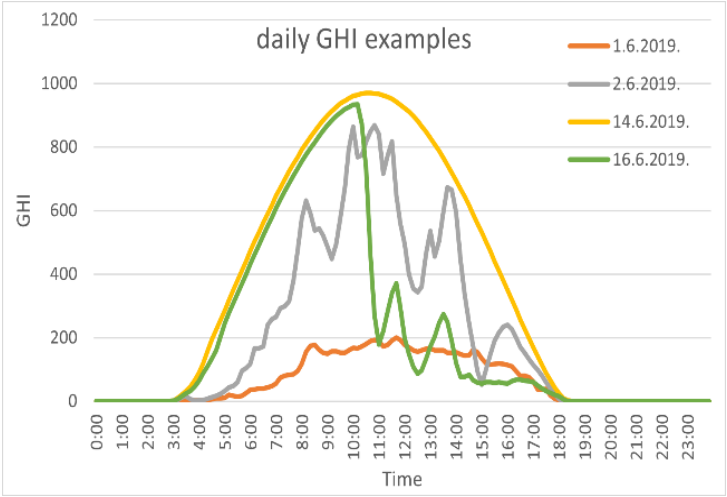

Figure 1. Example days of GHI

Preliminary Analysis of Attributes and the Corellation Between Them

The data set that enabled prediction of GHI contains 15 attributes, which are shown in Table 1, also they are sorted and numbered by relevance to the GHI prediction.

The Zenith Angle is most important for the prediction of GHI, it is the angle between the normal of the Earth's surface and the Sun, as it is shown in Figure 2.

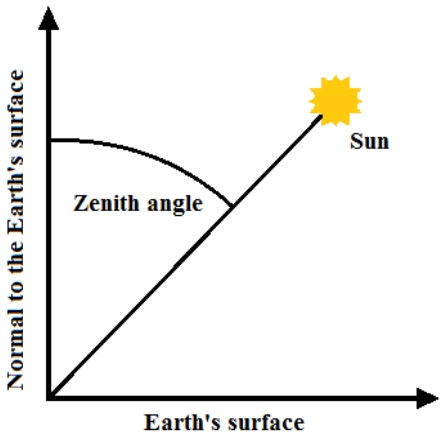

Figure 2. Zenith angle

Relative Humidity is estimated humidity for observed location, and it is presented in percentages (%).

Cloud opacity is the value of how opaque clouds are for sunlight in the observed region, and it is also presented in percentages (%).

Dew Point Temperature is the temperature of air for which air is saturated with water vapor and in contact with colder items it starts to condense on the item's surface [8].
Precipitable Water is the amount of water concentrated in the column of clouds that extends above some surface on the Earth, potentially available for precipitation [9]. Measured with units of kilograms on meter square (kg/m²). Albedo Daily is average daylight surface reflectivity, presented with a value between 0 and 1, where 0 is total absorption and 1 is total reflection.

To analyze the data and their relationship, Pearson correlation coefficient was used, where the value ±1 represents perfect positive (+1) and negative relationship (-1), from ±0.5 to ±1 represents a strong relationship, from ±0.30 to ±0.49 represents a medium relationship, and below ±0.29 represent a small relationship. Pearson correlation coefficient for two sets of values, x and y, is given by the formula:

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \sum(y - \overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of the two arrays of values.

Between the attributes in Table 1, high positive degree of correlation was found between the following attributes:

- Participle water and Dew point temperature,
- Air temperature with Dew point,
- Air temperature with precipitation of water,
- Snow depth with Albedo daily,
- Zenith with Relative Humidity,
- GHI with Air temperature.

A high negative correlation degree was found between the following attributes:

- Zenith with GHI,
- Relative Humidity with GHI.

A moderate degree of correlation was between the following attributes:

- Relative Humidity with Cloud Opacity,
- Wind speed with Cloud Opacity.

According to the Pearson correlation coefficient, for GHI, the zenith angle is the most important attribute, because they have a high negative degree of relationship (-0.764), and the attribute Air temperature, having a high degree of relationship of 0.502.
The highest degree of correlation between attributes was found between Participle water and Dew point temperature being 0.892, meaning for an increase in Participle water Dew Point temperature increases accordingly and vice versa.

Also, there is a high degree of correlation between the Air temperature with Dew point temperature being 0.889, and precipitation of water is 0.788. With high temperature, water evaporates into the atmosphere, hence increasing the amount of water in the air, affecting dew point and precipitation of water attributes. Snow depth and Albedo Daily have a high degree of correlation of 0.621. The reason behind this is that the snow being of the white color is more reflective than the darker ground surface [10].

Zenith and Relative humidity correlating 0.563. As the Sun is positioned normal to the Earth's surface, the angle decreases. During that period of the day, the Sun irradiating the Earth's surface is the highest, hence increasing air temperature. As the temperature increases, in turn, the humidity decreases and vice versa. This is also shown by the fact that the correlation between Zenith and air temperature is -0.504, indicating an inverse correlation between the attributes. In other words, when the Zenith angle increases, the air temperature decreases and vice-versa.

Table 1. Attributes used - sorted relevance to GHI predictions

	Attributes
1.	Zenith Angle
2.	Relative Humidity
3.	Air Temperature
4.	Cloud Opacity
5.	Dew Point Temperature
6.	Precipitable of Water
7.	Time
8.	Albedo Daily
9.	Snow height
10.	Surface Pressure
11.	Wind Speed
12.	Wind Direction
13.	Month
14.	Day
15.	Year

4. Modeling and Results

After the extensive analysis and tune up of the data set, it is ready to be fed to the machine learning algorithms. For that experimental setup, a specialized machine learning software WEKA is used in order to predict solar irradiation.

WEKA stands for Waikato Environment for Knowledge Analysis, and it is free software developed at the University of Waikato in New Zealand [11]. WEKA provides usage of many different machine learning tools and techniques. Data filters, data visualization, data classifiers and attribute selection tools were utilized in this study. Classifiers are algorithms that perform the process of predicting the class attribute, in this project GHI, of a given data points. Because GHI is in a range of 0 to 1000, it can be predicted using regression algorithms. Table 2 shows the statistics of the GHI from the used data set, including Minimum value, Mean value, Maximum value, and Standard deviation [8].

Table 2. Summary

Statistics	**Value**
Minimum	0
Maximum	991
Standard deviation	153,74
Mean	240,97

The M5' tree algorithm is a tree-based model algorithm, which constructs trees that can have multivariate linear models. By comparing it to the regression trees algorithms, M5' algorithm learns efficiently and it is good in tasks including big data sets and a high number of attributes [12].

After extensive testing and attribute selection processes, a summary of the results of the M5' algorithm are presented in Table 3, including Correlation coefficient, mean absolute error, Root mean squared errors, Relative absolute error, Root relative squared error and the Total number of instances.

Table 3. Performance of predicting GHI

Cross Validation	
Correlation coefficient	0,9995
Mean absolute error	3,6793
Root mean squared error	7,3679
Relative absolute error	1,9705%
Root relative squared error	3,0576%
Total number of instances	733383

The mean value from Table 2 is an average value of GHI in the data set, and the Mean absolute error from Table 3 is the average value of the predicted GHI error, so from these two values and by comparing them it can be known how good predictions are. Besides numerical interpretation, the precision of the predicted test data can be graphically interpreted, which is shown in the following Figure 4.

The next interesting step is attribute selection, which are collected by repeating model algorithm and every time excluding one attribute in the specific order of the attribute relevance for GHI prediction, like it is shown in Table 1.

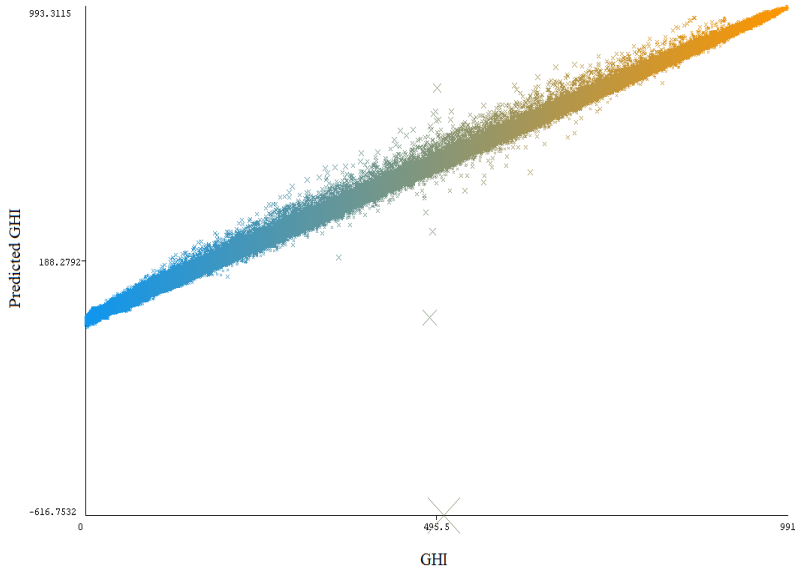

Figure 3. Scatter plot

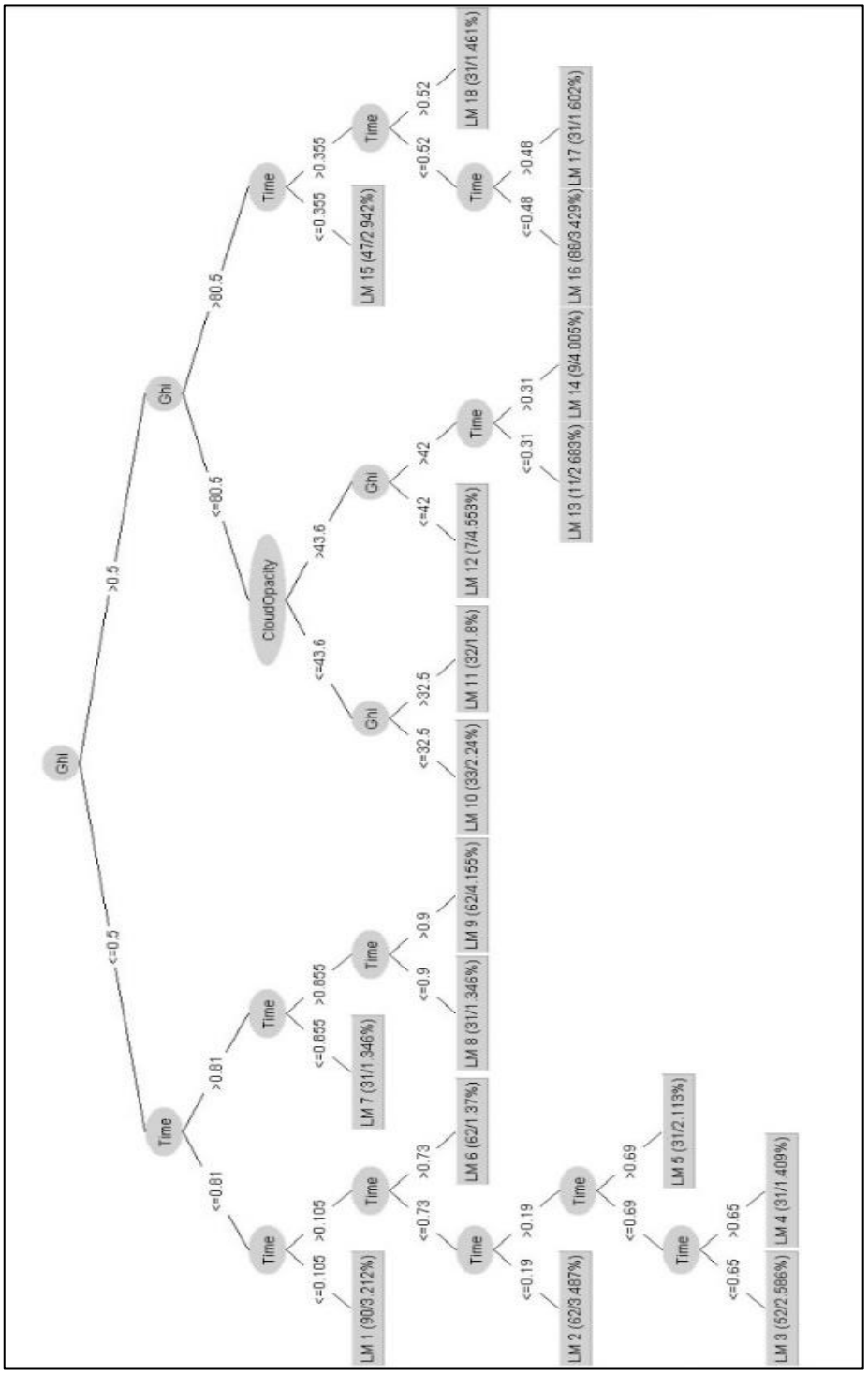

Figure 4. The tree of algorithm

The tree of our M5' with 1000 instances is shown in Figure 4. The results of this attribute selection are shown in Table 4.

Table 4. Attribute selection

Number of attributes	Attribute selection		
	Previously excluded attribute	***Correlation coefficient***	***Mean absolute error***
15		0,9995	3,6793
14	Year	0,9995	3,6904
13	Day	0,9995	3,6928
12	Month	0,9995	3,7509
11	Wind direction	0,9995	3,769
10	Wind speed	0,9995	3,7736
9	Surface pressure	0,9995	3,7776
8	Snow height	0,9995	3,8599
7	Albedo daily	0,9995	3,9416
6	Time	0,9995	4,0076
5	Precipitable of water	0,9993	4,6096
4	Dew point temperature	0,9992	4,7918
3	Cloud opacity	0,9281	44,8639
2	Air temperature	0,9231	47,1382
1	Relative humidity	0,8744	63,5234
0	Zenith	-0,003	186,7254

5. Conclusion

This study utilized the M5 regression tree for solar irradiation prediction. Azimuth angle was determined as the most important feature, year was the least important feature in the dataset. By excluding the least important feature, the correlation coefficient did not change. However, by its exclusion, the mean absolute error increased. Besides that, it also showed how solar energy sources and power grids with them can be easily upgraded in their operation. The proposed study has the potential to be even more precise and developed in the working, highly precise solar irradiance prediction tool, which could be used in some real-life applications.

References

1. A. M. P. O. A. V.-K. V. I. Mehrnoosh Torabi, "A Hybrid Machine Learning Approach For Daily Prediction of Solar Radiation," in Lecture Notes in Networks and Systems 53, 2018.

2. D. S. Alex Kim, "Predicting Solar Power Generation from Weather Data," Stanford University, 2019.
3. Srivastava, R., Tiwari, A. N., & Giri, V. K. (2019). Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. Heliyon, 5(10), e02692.
4. Guher, A. B., Tasdemir, S., & Yaniktepe, B. (2020). Effective Estimation of Hourly Global Solar Radiation Using Machine Learning Algorithms. International Journal of Photoenergy, 2020.
5. Keshtegar, B., Mert, C., & Kisi, O. (2018). Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM,
6. Meenal, R., & Selvakumar, A. I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renewable Energy, 121, 324-343.
7. "Solcast," Solcast, [Online]. Available: https://solcast.com. [Accessed 2020/2021].
8. Cassel, Encyclopedia of Soils in the Environment, North Carolina: North Carolina State University, Raleigh, NC, USA, 2005..
9. "Department of Atmospheric Sciences (DAS) of the University of Illinois," University of Illinois, 2010. [Online]. Available: http://ww2010.atmos.uiuc.edu/(Gh)/guides/maps/sfcobs/dwp.rxml.
10. Albedo and reflective properties of various types of snow and water | AMAP. (n.d.). AMAP. Retrieved May 10, 2021, from https://www.amap.no/documents/doc/albedo-and-reflective-propertis-of-various-types-of-snow-and-water/971
11. "waikato," University of Waikato, [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/ .
12. J. R. Quinlan, "Learning with continuous classes," Sydney, 2006.

Letter Recognition Using Machine Learning Algorithms

Merima Ćeranić, Samed Jukić
International Burch University,
Sarajevo, Bosnia and Herzegovina
merima.ceranic@stu.ibu.edu.ba

Original research

Abstract: *Optical character recognition represents the mechanical or electronic conversion of handwritten, typed or printed images into coded text. Optical character recognition is widely used as a form of data entry from records that have been printed, and it can include invoices, bank statements, passports and many more. In the research, Optical character recognition reads data from the Re-Captcha dataset of images, converts them into strings, and these strings are used for testing, training and calculating prediction accuracy. The methodologies used are Convolutional neural network and Recurrent neural network. The convolutional neural network consist of neurons that receive data and group them according to similarity. A recurrent neural network cycle can be created between the connections of nodes, allowing the output from nodes to influence the subsequent input to other nodes. For data were used Re-Captcha images, and for the prediction of characters from images was used TensorFlow with Keras. The best results that are produced can be compared between first and last result, where the loss for first result was 20.63 and value loss was 16.45, while last result has loss of 0.56 and value loss of 2.96.*

Keywords: Keras, OCR, Re-Captcha, Tensorflow.

1. Introduction

Institutions that focus on finance, such as banks, are involved in the process of creating the latest records for their clients or may even include the conclusion of new deals, which creates many records that are in paper form. Thus, documents become numerous, and everything is in the form of paper, and they can become a big concern for banks. These documents are crucial, and the ideal solution is to digitize these data, and using a data entry service is the ideal solution for this problem. Today's security is in digital documents, which means simple storage and very fast retrieval of documents. Banking institutions must simplify organizational processes in order to provide their users with the best possible services. In this way, the search time for all documents used by banks, for example, would be reduced. The technology that is unique to this is Optical character recognition (OCR), which is used by banks as part of the extraction of huge amounts of information. With OCR, banks can process, evaluate and monitor payments including huge amounts of data about their customers. These data are most often personal data or security data.

OCR is a technology that can extract all text from images, documents or scanned files. It enables banks to minimize human error, and to save time and effort while simultaneously improving the user experience. Banks should properly authenticate customers for routine or banking transactions, account opening and numerous other functions. For example, with the help of OCR or machine learning, banks can extract data in real time from passports or other documents. In this way, they can quickly identify clients before transferring money or opening a new bank account. OCR provides a Software development kit (SDK) that includes personal document understanding, data identification, and data validation. It can check whether the signature on the personal document matches the signature of a real person. Some of the tasks of Optical character recognition are mentioned below, where are listed some of the most important usings. Task of Optical character recognition are:

- The density of the text on the written station represents dense text. So, for example, every day we can notice a STOP sign on the streets, where the text is scarce.
- Text structure: Generally, the text on the station is structured in strong lines, while the example works in external conditions - the wilderness, the text can be scattered anywhere.
- Fonts: Written letters are darker, while printed fonts are lighter because they are more structured than written letters.

- Type of sign: We can distinguish many languages, so the text can also differ. One example can be numbers, where we immediately have a difference in house numbers.
- Artifacts
- Location: some tasks include a centralized text, while sometimes the text can be scattered in random places [1].

Also, in addition to the previously mentioned ways of using OCR, in everyday life we encounter applications that require additional authentication whether the end user is a robot or not. In those cases, the most common display of authentication is expressed in the form of Re-Captcha images. The paper shows how to use OCR for Re-Captcha images. Furthermore, OCR can help to read the data from the image, and further this data can be converted into audio recordings. If Re-Captcha images are converted into audio recordings, these recordings can be used by blind people.

2. Literature Review

Character recognition has become an integral part of computer analysis and vision. Several corporations are working on improving the technique and the current situation. Algorithms that can recognize notes and handwritten numbers are being actively developed by several industries. In the era of digitization, editing, indexing, finding and storing information in digital documents is much easier than spending a lot of time flipping through printed documents. Searching for data in a non-digital document is not only time-consuming, but there is a very likely possibility that some information will be missed while manually searching a document. Every day computers are getting better and better at doing tasks that people thought only they could do.

Related work

In the research by Karishma Tyagi and Vedant Rastogi, it can be observed that OCR recognizes any multimedia content like videos and images. We can use character positioning, image processing, neural network to solve the problem of image recognition in text. According to the research of these two authors, there are approaches for identifying links:

- *HMM approach*
  The Hidden Mark model is a stochastic two-step process. An established stochastic process that can be observed as another stochastic process that produces a series of

observations, but is not visible.

- *A neural network approach*
  Recognition of registrar characters on license plates plays a significant role in the optical recognition system. This can be directly related to the success or failure of system recognition.

- *Normalization of character*
  The necessary step is to frame the letters, characters or numbers to some standard size. Character normalization to one fixed size can be performed to simplify the task of optical character recognition.

- *Correlation method for recognizing the lower sign*

- *Pre-processing*
  This step involves converting images to grayscale. Then the image is converted into a binary image. This process is also known as the process of image digitization. It is important to note here that the picture may have some defects or difficulties. The result may be some unnecessary details that are present in the image.

- *Segmentation*
  In this step, the position of the object is learned. The size of the image can be expressed according to the size of the template [2].

In the article by Konica Minolta, How optical character recognition works, OCR allows users to convert scanned images into text and to convert paper documents into digitized documents. Digitizing documents helps reduce the amount of physical space required to store documents. Also, digitizing documents can reduce the risk of lost or incorrectly archived documents and, in many cases, eliminate the need for manual processing of documents. Manual processing of documents in most cases leads to errors. According to Konica Minolta, OCR analyzes the patterns of light and darkness that makeup letters and numbers to turn a scanned image into text. OCR recognizes characters in a variety of fonts. Early OCR systems were designed to work with one specific font, which was created specifically for this purpose. Today, modern OCR systems can even recognize people's handwriting. The technology that recognizes human handwriting is called intelligent character recognition (ICR). OCR programs work on the principle of recognizing text character by character. It can also check errors during the process at the end of the process.

OCR as a technology has existed since the late 1920s. Today, OCR can convert large documents, where only a few errors can occur. It is important to mention that there are six key ways OCR helps businesses [3]:

- *Automate workflows*
  Businesses that work with large amounts of paperwork can save time, and thus can increase productivity through scanning.

- *Turn read-only files into editable text*
  OCR allows users to read PDF documents that can be later edited, and used in other documents, but can also be searched.

- *Create audible files*
  It saves time spent reading long and complex documents. It enables the conversion of documents into a document that the user can listen to (while going to work), and it is considered that in these situations the user becomes more productive.

- *Translation of foreign documents*
  Some OCR solutions can convert documents into more than 180 foreign languages.

- *Manage forms and questionnaires*

- *Achieve faster, more accurate data entry*

In an article by Nitin Ramesh, Aksha Srivastava and K. Deeba, Improving Optical Character Recognition Techniques, it is stated that there are two types of character recognition, printed and handwritten. In the printed character recognition type, OCR searches for written text and reviews it one by one. On the other hand, Intelligent character recognition (ICR) can also work with text that is handwritten [4]. For the handwritten type of character recognition, the offline way to recognize is static document processing, while the online version is much more advanced and uses handwriting motion analysis. The most commonly used algorithm for learning patterns, the online mode allows us to record the movement, that is, the order in which the segments are drawn and what is their direction of movement [5].

3. Methods

In this part of the paper, the methods used will be described, as well as a database. The method will be presented as a flowchart, where each process will be described step by step.

Machine Learning Methods

The research includes Recurrent neural network (RNN) and Convolutional neural network (CNN) as machine learning methods. A convolutional neural network (CNN) is a machine learning algorithm used for image processing, recognition and classification for face detection or object identification. It consists of neurons that receive data that later assign importance to them and group them according to similarity. It is also called "ConvNet", in order to make accurate predictions it can look at the surroundings of the object. They will look at smaller parts or letters, rather than looking at the whole picture to determine features [6]. The recurrent neural network (RNN) represents artificial neural networks, where a cycle can be created between the connections of nodes, allowing the output from some nodes to influence the subsequent input to other nodes. RNN can use internal state to process variable-length nodes. The term "recursive neural network" will be used to denote the class of networks with infinite impulse response, while the previously mentioned CNN refers to the class of finite impulse responses. Both classes represent temporal dynamic behavior. A recursive network with a finite impulse is a directed acyclic graph that can be replaced by a neural network with strict forward connection. A recursive network with infinite momentum is a directed cyclic graph that cannot be wrapped [7].

Importing Libraries

The first step was to import appropriate libraries that will be required for the needs of the project. Some of the libraries that were using are numpy, matplotlib, tensorflow. Numpy is Python library that provides multidimensional array object. The elements in NumPy are all required to be the same data type. Matplotlib is data visualization and graphical plotting library for Python ant its numerical extensions NumPy. Tensorflow is open source library for numerical computation that makes machine learning and development neural network faster and easier.

Collecting Data

For this purpose is decided to use Re-Captcha images, which are collected from the internet. The dataset that is used in the project contains 1040 Re-Captcha PNG files. The label for each sample is a string.

Processing

Data processing is the method of transforming data into a graspable format. Collecting data is usually incomplete, noisy, inconsistent, and redundant. Processing data is an important step to reinforce data effectiveness. This step is done in the Jupyter notebook and includes mapping characters to integers, and then again, mapping integers back to original characters.

Create Dataset Objects

In this stage were created training and validation datasets. This was done by using TensorFlow functions.

Visualize Data

Visualization is important step because through this step can be seen with what data we are working, and what is real output of it. In this part of the work was used function from matplotlib.

Build a Model

In this stage computing the training time loss value and adding it to layer using self.add_loss() function was done. In this part of the research was used functions from tensorflow.keras – keras, and was used keras from tensorflow.

Train a Model

Training is a step that follows the steps of Building a model. In this step, we could see what is value for loss and what is value for value loss. In this part of the research was used function from keras. Import step of keras is mentioned in *G* step.

Printing the Interface – Printing the Results

The last step that we were using in our research was to print the results. Through this step we can see how is model trained. *Figure 1* presents stages that were used for research process.

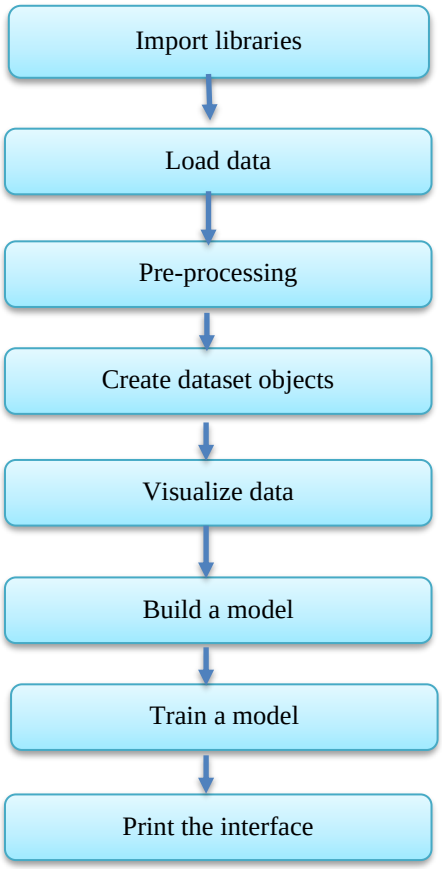

Figure 1: Diagram – Stages (B – I) needed for research

4. Results

To calculate the loss of value and calculate the loss, the model.fit was used in the work. Model.fit and everything that works smoothly can be used to calculate loss of value and loss. Model.fit is a customized training algorithm. Progressive complexity discovery is the basic tenet of Keras. How well a machine learning model generalizes data similar to what it was trained with is the basic story of a model measure. A good model fit refers to a model that accurately approximates the output when it has invisible inputs. Setting up readers to adjust parameters in the model to improve accuracy involves running the algorithm on data for which the target variable is known to produce the machine learning model. Model results are compared to actual values of the target variable to determine accuracy. The next step involves adjusting the standard parameters of the algorithm in order to reduce the error rate and make the model as precise as possible in determining the relationship between the target variable and the features. Until the model finds the optimal prediction parameter with significant accuracy, this process is repeated. *Table 1* shows the first five and the last five epochs of the training model where the time of execution, loss of value and loss can be seen. In the results can be seen the progress of the model. At the first epoch, loss was

20.6370 and value loss was 16.6513, while on the last epoch loss was 0.5616 and value loss was 2.9697. How well model is trained can be seen with comparison of loss 16.6513 and 0.5616, and value loss 16.6513 and 2.9697.

Table 1. Epochs – time – loss and values loss for first five epochs

Epoch n/100	Time	Loss	Value loss
1/100	43s 402ms/step	20.6370	16.4513
2/100	20s 339ms/step	16.3671	16.4555
3/100	20s 332ms/step	16.3534	16.4488
4/100	19s 325ms/step	16.3467	16.4365
5/100	22s 375ms/step	16.3346	16.4167
96/100	2s 29ms/step	0.6674	3.1233
97/100	2s 29ms/step	0.6018	2.8405
98/100	2s 28ms/step	0.6322	2.832
99/100	2s 29ms/step	0.5889	2.8786
100/100	2s 28ms/step	0.5616	2.9697

Figure 2 presents a prediction of a model. There can be seen sixteen examples and what is prediction for each of them. Sixteen examples present visual presentations of results, so it can be easily seen that for last example in forth column is 6p2ge.

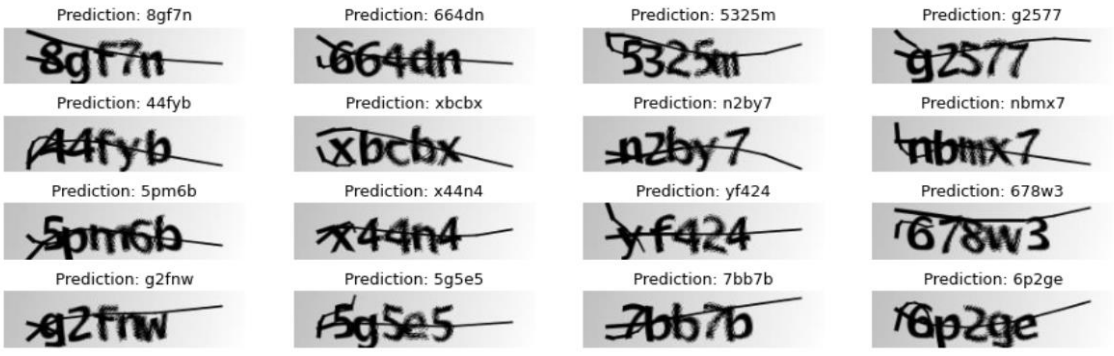

Figure 1. Prediction of a model – Visualization for 12 examples

5. Discussion

By following TensorFlow using Keras, has been seen how can be calculated accuracy of reading letters from Re-Captcha images. Analysis of data in the research is done by using machine learning (ML) tools. By looking at accuracy, that is presented above images/examples (*Figure 2*), it can be seen that accuracy is correct. Most letters from images are correct, which is pretty amazing. It can be concluded that model is trained well when the comparison between first and last epoch is made. For example, the results for first epoch were 20.6370 for loss and 16.4513 for value loss, while for the last epoch the results were better 0.5616 for loss and 2.9697 for value loss. Optical character recognition is useful for cases of scanning documents and transforming them into digital form. Besides, optical character recognition is useful for scanning documents, such as passports, transforming documents into audio format, etc. For example, if we take a look at one of the images (*Figure 3*) that was used in the paper, it can be seen that the prediction was pretty correct.

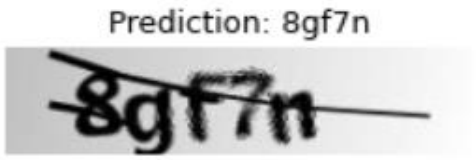

Figure 2. Prediction for single PNG file

In the research by Ondrej Bostik and Jan Klecka, Recognition of Captcha Characters by Supervised Machine Learning Algorithms [8], they reveal that all used algorithms can classify objects into the first class with a 99% success rate. Computational costs are the main difference in algorithms. According to the research, Pettr's neutron network proved to be the best algorithm. It has good precision and good computational costs. Feed-forward is the name for an inappropriately used neural network that was not optimized for pattern recognition. It had great difficulties during the learning phase when achieving correct performance.

In the research conducted by Ashish Renjan, Varun Nagesh Jolly Behra and Mothahar Reza, OCR using Computer Vision and Machine Learning [9], OCR basics were studied. Preprocessing is an important step in OCR systems. The authors state that if the input image contains tabular data, the processing becomes more complex if the tabular structure is to be

preserved. In conclusion, contours can help in extracting tabular data. The extraction succeeds if the borders are visible, but fails if the borders are not visible.

In the research by Jangid and Srivastava's from 2019, they used deep learning techniques to improve existing OCR approaches for recognizing Chinese capital letters [10]. The deeper the number of neural network layers and the more parameters, the more accurate the results. Accurate results mean that more computer resources are used. Removing data that is not essential for research is of enormous importance. This can be achieved by identifying the most connected neurons using the Average Percentage of Zeros algorithm, and removing some unnecessary network neurons and keeping the weight parameters that are key to reducing the network parameters in order to reduce the reasonable complexity of the model. The accuracy was reduced by 1.26%, but a 96.5% net weight reduction was achieved.

In the research of Khaled S. Younis and Abdullah A. Alkhateeb, A New Implementation of Deep Neural Networks for Optical Character Recognition and Face Recognition [11], using TensorFlow to classify the ubiquitous MNIST dataset, they designed a multilayer neural network. The accuracy was 98.48%.

7. Conclusion

Optical character recognition helps users to recognize numbers, letters and written characters. It can convert images and scanned documents into electronic data. This technology is at the center of a growing trend when it comes to workflow modernization. Artificial intelligence (AI) opens the door to many possibilities, and OCR can be used for many purposes. If we pay attention to the work of Konica Minolta and make a comparison with this work, we can conclude that they have the same goal. In working with Konica Minolta, the main goal was to document data and how to preserve it from loss. During my research, the goal was to see how accurately we can read from images, how accurate the prediction is for each image.

According to research by Himini Kohli, Jyoti Agarwal and Manoj Kumar [12], printed characters are easy to recognize because they have a defined size and shape. OCR faces the difficulties of handwriting, as each individual has a different handwriting. To solve the problem, the OpenCV technique is used in the research, which has a focus on testing and training the model. From the research, 99.5% training accuracy and 99% testing accuracy were achieved.

Reference

1. G. Shperber, “A gentle introduction to OCR”, 2018
2. K.Tyagi, V. Rastogi, “Survey on Character Recognition using OCR Techniques”, 2014
3. K. Minolta, “How optical character recognition works”, 2018
4. N. Ramesh, A. Srivastava, K. Deeba, “Improving Optical Character Recognition Techniques”, 2018
5. D. G.Pelli, C W.Burns, B. Farell and D. C. Moore-Page, “Feature detection and letter identification”, 2006
6. S. Saha, “A Comprehensive Guide to Convolutional Neural Networks - the ELI5 way”, 2018
7. N. Laskowski, “Recurrent neural networks”, 2021
8. O. Boštík, J Klečka, “Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms”, 2018
9. A. Ranjan, V. Behera, M. Reza, “OCR Using Computer Vision and Machine Learning”, 2021
10. Yin Y, Zhang W, Hong S, Yang J, Xiong J, Gui G., “Deep Learning-Aided OCR Techniques for Chinese Uppercase Characters in the Application of Internet of Things”, 2019
11. K. Younis and A. A. Alkhateeb, “A New Implementation of Deep Neural Networks for Optical Character Recognition and Face Recognition”, 2017
12. H. Kohli, J. Agarwal and M. Kumar, “An improved method for text detection using Adam optimization algorithm”, 2022

Call for new submission

Background

JONSAE provides a platform for the researchers, academicians, professionals, practitioners and students to impart and share knowledge in the form of high quality empirical and theoretical research papers. The journal covers all areas of Genetics and Bioengineering, Electrical and Electronics Engineering, Information Technology, Architecture, Applied Mathematics, Computer Sciences and Civil Engineering.

Submission and review process

All submissions should be made to email jonsae@ibu.edu.ba.

Submissions must adhere to the format and style Guidelines for JONSAE articles available on the journal's web page.

The official referencing style is APA.

Submissions will be subject to an initial screening by our editors and papers that fall outside the scope or which are considered unlikely to be suitable for the JONSAE issue will be desk rejected.

Accepted papers will undergo a typical double-blind review process.

Types of submission

We welcome high-quality submissions which advance our knowledge on the above mentioned topics. We do not favors any special theoretical perspectives or methodological approaches. The types of acceptable submissions include, but are not limited to:

Theoretical and empirical papers
Literature reviews
Practice reviews
Qualitative, quantitative, mixed-methods research
Experimental research
Single, multiple, large-sample case studies

For any questions, please contact us at jonsae@ibu.edu.ba or publication.office@ibu.edu.ba.