CHAPTER 2

LITERATURE REVIEW

This chapter introduces Digital Quran and summarizes the architecture of the Digital Quran model. It also presents related works that address issues on the Digital Quran model. The research gaps are identified and analyzed to justify this thesis. Figure 2.1 outlines the structure of this chapter.

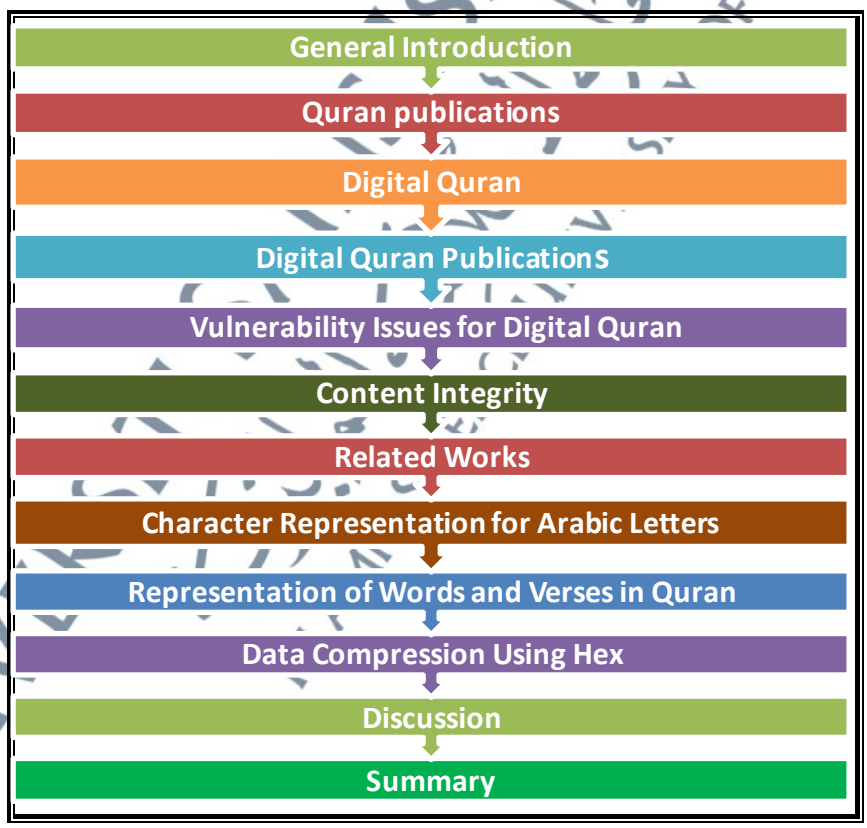


Figure 2.1: Outlines the Structure of Chapter 2

2.1 Arabic Language Foundation

The Arabic language is written using some 28 letters, where 16 of them have one dot, two or three dots. Arabic is written from right to left, and numerous kinds of fonts exist, with letters changing their shape according to the place they occur in the text.

The Quran contains 114 chapters, 30 juzu, 6236 verses, 77797 words and 330,709 letters ((https://qurananalysis.com/analysis/basic-statistics.php). As the number of words is considerably high, storage optimisation and safe searching time are essential. Accordingly, the current research builds on the existing literature and aims to develop a model that can further optimize storage usage for the digital Quran. This further has been explained in the previous chapter (see chapter 1). It is important to note that the current gaps within the literature, according to several studies, are the storage optimization aspect (Almazrooie et al., 2020; Hakak et al., 2017; Saada & Zhang, 2015; Mouratidis et al., 2013). In addition, the Arabic language requires further analysis and evaluation regarding applications and UTF measures (Almazrooie et al.,2020). Within the extant literature on the subject that Chinese, Hindi and Arabic languages have been addressed by several techniques to be represented (Law & Chan, 1996). For instance, Chinese characters are approximately 20,000, with 6.700 commonly used (Law & Chan, 1996). There are compound words shaped by these characters that can vary in length such as "海上" and "上海" as above (上) and sea (海). The word "海上" translates into above the sea. However, t上海" means Shanghai (Almazrooie et al.,2020).

After English and Chinese, the Hindi language comes third in the context of a Unicode function being retrieved from the web (Tripathi, 2012). This has been linked to a lack of understanding of the language and how the Hindi language is presented (Tripathi, 2012). Limited literal matched patterns also complicate good algorithms (Sharma et al., 2012). Relevant to the context of this research, studies have included and investigated the Arabic language regarding its modern standard form. This has created a challenge for establishing Quranic Arabic (an ancient form of the Arabic language). However, according to several studies, the Quranic Arabic has significance for Muslims worldwide and is and not neglected (AlMaayah et al., 2014). In their study, a model was introduced that could decrease stop word, stemming and POS tagging through grouping words of the same meaning in speech parts. Other studies address the importance and challenges of the digital Quran. It has been established that transforming Quranic words in Arabic requires further optimization, analysis, and model developments to ease the app's usability for all users across all devices (Mouratidis et al., 2013).

A recent study noted that data integrity was emphasized and focused on information integrity checks (Almazrooie et al., 2020). The cryptographic hash function was used for the integration of transmitted data. Furthermore, they use a single compression technique to manipulate data during run time. This method uses two bytes in Unicode UTF-8 for Arabic characters as a set. Their findings revealed the size of the hash tables to be relatively minor (6.55 fold and 10.48 fold) than the original copy. While their study mainly focuses on data integration (integrity verification model), they suggest that future studies further propose models that can optimise storage usage. Additionally, they suggest

that future studies use various approaches to understand better the challenges and the issue at hand by specifically addressing compression methods (Almazrooie et al., 2020). The current research follows scholars' recent and relevant work in the literature to justify its conduct. In addition, this research includes DQM to assess the integrity check measure.

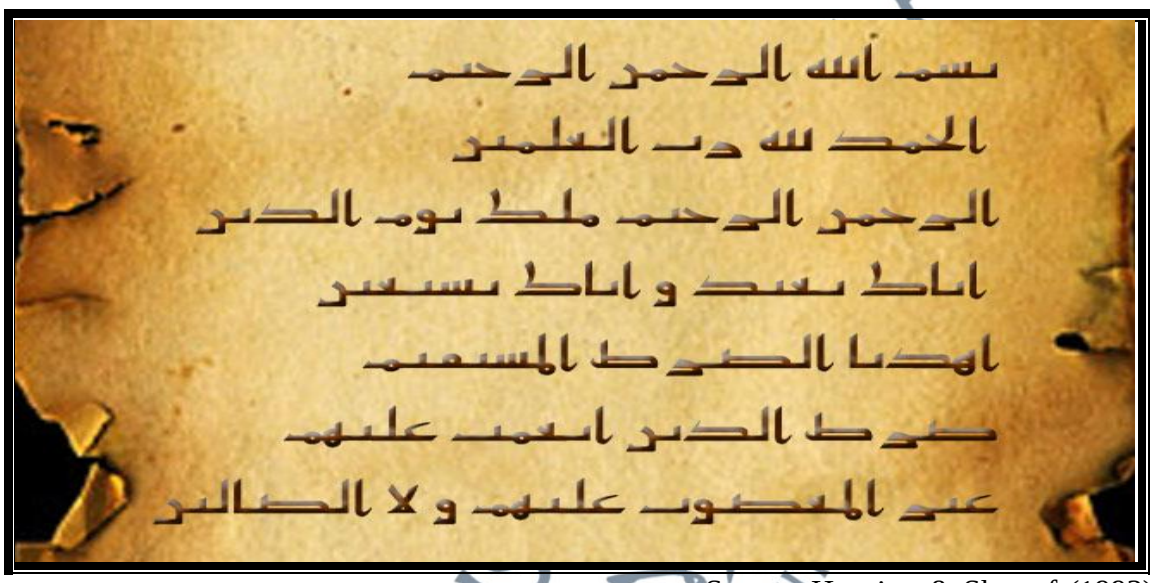
2.2 Quran Publications

The traditional narrative states that Profit Muhammed SAW had companions, and as scribes, they recorded the revelations in writing. They later brought together the Quran and wrote it down soon after his death. They also memorised parts of the Quran (Donner, 2006; Campo, 2009).

He ordered Hafsa, the daughter of Caliph Umar, to copy the Quran. Consequently, others were ordered and came up with their codices. Their manuscripts had a Quraish tone. They include Zaid ibn Thabit, Abdullah bin Zubair and Abdul Rahman bin Harith bin Hisham (Sadeghi & Bergmann, 2010). Caliph Uthman chose to set a new version, now called the Uthman's codex, which became the archetype of today's Quran. This was due to the codices' variations. Currently, different readings differing minor in meaning do exist (Donner, 2006).

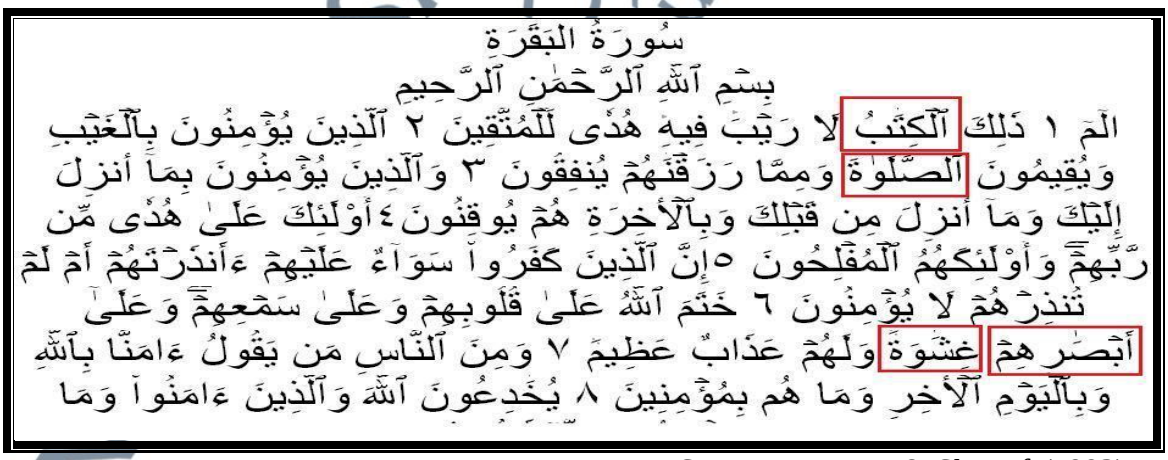
After Prophet Muhammad's death, the first Quran was compiled by Caliph Abu Bakar, the Caliph Umar completed by Caliph Uthman's time. He later established the Uthman's codex as the standard, th; thean was translated into many languages and manuscripts. The well-known are rasm Imlai and rasm Uthman (Rubin, 1998). Figure 2.2

shows Surat Al-Fatiha in Uthman typing without Dots. Figure 2.3 represent rasm Uthman Surat Al Baqarah, and Figure 2.4 for rasm imlai Surat Al Baqarah; the old Uthmani typing did not have dots.



. Source: Hawting & Shareef (1993)

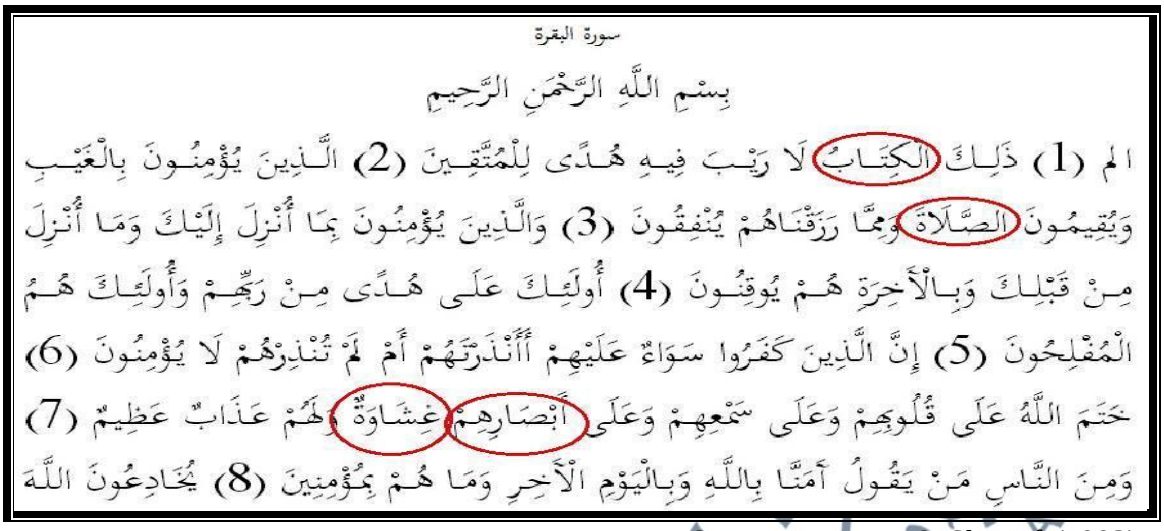
Figure 2.2: Surat Al-Fatiha without Dots Rasm Uthmani



سُورَةُ البَقَرَةِ
بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ
الٓمٓ ١ ذَٰلِكَ ٱلۡكِتَٰبُ لَا رَيۡبَۛ فِيهِۛ هُدٗى لِّلۡمُتَّقِينَ ٢ ٱلَّذِينَ يُؤۡمِنُونَ بِٱلۡغَيۡبِ
وَيُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقۡنَٰهُمۡ يُنفِقُونَ ٣ وَٱلَّذِينَ يُؤۡمِنُونَ بِمَآ أُنزِلَ
إِلَيۡكَ وَمَآ أُنزِلَ مِن قَبۡلِكَ وَبِٱلۡأٓخِرَةِ هُمۡ يُوقِنُونَ ٤ أُوْلَٰٓئِكَ عَلَىٰ هُدٗى مِّن
رَّبِّهِمۡۖ وَأُوْلَٰٓئِكَ هُمُ ٱلۡمُفۡلِحُونَ ٥ إِنَّ ٱلَّذِينَ كَفَرُواْ سَوَآءٌ عَلَيۡهِمۡ ءَأَنذَرۡتَهُمۡ أَمۡ لَمۡ
تُنذِرۡهُمۡ لَا يُؤۡمِنُونَ ٦ خَتَمَ ٱللَّهُ عَلَىٰ قُلُوبِهِمۡ وَعَلَىٰ سَمۡعِهِمۡۖ وَعَلَىٰٓ
أَبۡصَٰرِهِمۡ غِشَٰوَةٞ وَلَهُمۡ عَذَابٌ عَظِيمٞ ٧ وَمِنَ ٱلنَّاسِ مَن يَقُولُ ءَامَنَّا بِٱللَّهِ
وَبِٱلۡيَوۡمِ ٱلۡأٓخِرِ وَمَا هُم بِمُؤۡمِنِينَ ٨ يُخَٰدِعُونَ ٱللَّهَ وَٱلَّذِينَ ءَامَنُواْ وَمَا

Source: Hawting & Shareef (1993)

Figure 2.3: Rasm Uthmani for Surah Al-Baqarah.


سورة البقرة

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

الم (1) ذَلِكَ الْكِتَابُ لَا رَيْبَ فِيهِ هُدًى لِلْمُتَّقِينَ (2) الَّذِينَ يُؤْمِنُونَ بِالْغَيْبِ
وَيُقِيمُونَ الصَّلَاةَ وَمِمَّا رَزَقْنَاهُمْ يُنْفِقُونَ (3) وَالَّذِينَ يُؤْمِنُونَ بِمَا أُنْزِلَ إِلَيْكَ وَمَا أُنْزِلَ
مِنْ قَبْلِكَ وَبِالْآخِرَةِ هُمْ يُوقِنُونَ (4) أُولَئِكَ عَلَى هُدًى مِنْ رَبِّهِمْ وَأُولَئِكَ هُمُ
الْمُفْلِحُونَ (5) إِنَّ الَّذِينَ كَفَرُوا سَوَاءٌ عَلَيْهِمْ أَأَنْذَرْتَهُمْ أَمْ لَمْ تُنْذِرْهُمْ لَا يُؤْمِنُونَ (6)
خَتَمَ اللَّهُ عَلَى قُلُوبِهِمْ وَعَلَى سَمْعِهِمْ وَعَلَى أَبْصَارِهِمْ غِشَاوَةٌ وَلَهُمْ عَذَابٌ عَظِيمٌ (7)
وَمِنَ النَّاسِ مَنْ يَقُولُ آمَنَّا بِاللَّهِ وَبِالْيَوْمِ الْآخِرِ وَمَا هُمْ بِمُؤْمِنِينَ (8) يُخَادِعُونَ اللَّهَ

Source: Hawting & Shareef (1993)

Figure 2.4: Rasm Imlai for Surah Al-Baqarah.

Later, Imlai manuscripts came after Uthmani manuscripts for the holy Quran. They do have some differences. In the Uthmani manuscript, there exists a short form for words; for instance, the word of al Kitab, the book (ٱلْكِتَٰبُ), whereas in imlai is written with an additional letter (ٱلْكِتَابُ), another example the word pray (الصلاه, الصلوت) imlai and Uthman respectively; these two kinds of Quran fonts make the representation of the words and verses harder in digital Quran.

All over the world, Muslims consider Arabic the essential language. They view it as such because the Holy Quran was given as a miracle in the language. The language to meet the Ara and Islamic civilisation's needs at its peak of prosperity (Kanaan & Wedyan, 2006).

The Arabic language is made up of 28 letters. Sixteen letters have either a dot two dots or three them. The writings go from right to left. The letters can be written using

lotmanyailable fonts, including Tahoma, Akufi, and Andalus (Khafajeh et al., 2010), as presented in Table 2.1, which lists examples of Arabic fonts with their font names.

Table 2.1: Examples of Arabic Fonts and Arabic Fonts Names

Font	Example
Andalus	A : نقدّم في هذا البحث قاعدة بياناتٍ لكلماتٍ عربية
M Unicode Sara	B : نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
Simplified Arabic	C : **نقدّم في هذا البحث قاعدة بيانات لكلمات عربية**
DecoType Naskh	D : تقدم في هذا البحث قاعدة بيانات للكلمات عربية
Diwani Letter	E : تقدم في هذا البحث قاعدة بيانات لكلمات عربية
Advertising Bold	F : نـقـدم فـي هـذا الـبـحـث قـاعـدة بـيـانـات لـكـلـمـات عـربـيـة
Arabic Transparent	G : نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
Tahoma	H : نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
Traditional Arabic	I : نقدّم في هذا البحث قاعدة بيانات لكلمات عربية
DecoType Thuluth	J : نقدّم في هذا البحث قاعدة بيانات لكلمات عربية

Source: Hawting & Shareef (1993)

Today, smart mobile devices, personal computers, and tablet applications may be installed with digital Quran. However, searching, translating and explication can be improved, for example, requirements by the Quran user besides reciting it (Foda et al., 2013).

2.3 Digital Quran

Over the years, numerous applications have been developed to aid the user in accessing the Quran both online and offline. The developers have considered text searching and playing the digital Quran recitations. For those wishing to memorize the Quran or read its explanations (known as Tafseer), the internet has remained a valuable

tool among Muslims in learning the digital Quran. With time, there has been the development of better Islamic websites enhanced specifically for digital Quran learning.

In the literature on the subject, it has been reported that there has been a constant growth in the usage of the digital Quran since its first digital copy in 2007 (Hilmi et al., 2013; Mobile Holy Quran, 2007; Almazrooie et al., 2020). While the first version is image-based (a copy of the original book), there have been two forms of digital Quran (image or text-based copies). There are pros and cons to each of these types; notably, Quranic applications designed in recent years do not use either of the formats mentioned earlier (Almazrooie et al., 2020). In recent years the digital format of the Quran has extended to be pdf (both image and text), text files, applications, e-books, raw data (Unicode) and more. This is while the sources that publish digital Quran are well-established organisations with sponsorship from governmental bodies (e.g. King Fahd Glorious Quran Printing Complex (2018). It is important to note that some copies or digital forms of the Quran are the results of volunteer work, such, as The Nobel Quran (2016), The Quranic Arabic Corpus by Dukes (2009), and Tanzil (2007). It can be said that the collective aim is to provide an authenticated format of the Quran that is easy to use and convenient across various platforms and devices.

Digital Quran has seen an increase in usage worldwide, which has led to the need for software and application development which can foster knowledge while maintaining the authenticity of retrieved information from the textbook. Accordingly, several researchers have addressed Quran and its relevant studies concerning technology (e.g. Adhoni et al., 2013a, 2013b), where mobile-friendly Quran applications and cloud-based

programming for Quranic applications have been studied. In addition, the semantic method for query translation has assessed Quranic applications, which examines cross-language information retrieval (CLIR) (Yunus et al., 2013). Arabic, Malay, or English query translations were examined, and reports showed variations among findings. In this sense, applications could vary from 638Mb to 79kb in size (Khan, & Alginahi, 2013).

The application above is in progress parallel to advancements in Multimedia technologies (Karkar et al., 2015). Several sources are used (e.g. websites, Quran portals, or smartphones) for learning Quran (Adhoni & Siddiqi, 2013). According to a report by Hakan et al. (2017), more than 70% of participants used the internet to refer to or seek a particular Quranic verse or hadith, while over 50% preferred a soft copy on mobile devices. This shows that the number of users of the digital Quran is relatively increasing as more Muslims can use smartphones and other technologies (e.g. internet) to seek a digital version of the Quran and hadith. This is supported by the extant literature and is a gap on which researchers can further conduct studies. Hence, the current research follows its structure accordingly (Adesina et al., 2010).

Text watermarking feature is categorized as both linguistic and nonlinguistic (Adesina et al., 2010). Linguistic techniques manipulate a document's lexical, syntactic and semantic properties. This is while maintaining the original meaning of the document. In contrast, nonlinguistic methods and techniques imply variation in texts through text attributions or embedded messages. Numerous approaches and techniques can be named in both disciplines: word-shift coding and feature/character-coding, natural-language-

based, and synonym substitutions or semantic-transformation techniques (language-dependent) (Adesina et al., 2010).

In light of what was noted, the current research addresses the implications of storage usage optimization. As noted earlier, users can benefit from decreased storage required for the app; additionally, the model in the current study aims to reduce the time it takes for character encoding by the usage of UTF-8 and sparse matrix, which follows the work of Diwakar et al. (2010), and Almazrooie et al. (2020).

The technology of the digital Quran often used to facilitate data input in the form of text is Optical Character Recognition (OCR) technology. OCR technology is the process of translating the character (image character) into text form by matching the pattern of characters per line with a pattern that has been stored in a database application.

Several studies have explored different areas related to the digital Quran. Ta'a et al. (2017) developed a web-based using PHP and MySQL database and studied the relationship between Quran and information technology in terms of searching for the classification of al Quran. Ahmad et al. (2016) explore a digital Quran for Malay Qur'an Readers that focuses on the search techniques of the Quran. Adhoni and Siddiqi (2013) built a digital Quran search API for learning al Quran, whereas Ouda (2015) built an "Intelligence System" also for the digital Quran.

2.4 Digital Quran Model Development

Following what was noted above, this study follows the work of Norman & Yasin (2013) in which they report that software developers do not have a clear and vivid procedure or KPI standards. This, in turn, yields a significant challenge for users as the

source material can be biased, w further decreasing the application's validity and reliability. In their study, common certainty management standards (Systems Security Engineering Capability Maturity Model – SSECMM) were used to examine the reliability of the online application. Additionally, they examined standard criteria definition and dimensions of certainty of SSE-CMM. They define certainty elements for application developers through a case study thaisre significant for the context of current research.

According to Norman and Yasin (2013), reliability issues surrounding digital Quran development are not significant. However, as the number of users grows alongside the number of online applications and platforms, the concern for reliability remains. This has been further noted by other studies, such as Alzamoorie et al. (2018, 2020). Under the work of Norman and Yasin (2013), SDLC has been deemed appropriate for software development. The current research recognizes various findings in the literature to provide and establish a thorough understanding of the matter at hand. Researchers show consensus regarding the fact that authenticity aspects in essence challenge Digital Quran Model Development. Thus, it is appropriate for the current research to address this issue.

As there is a lack of consensus on the DQM matter, it is essential to highlight that novice users risk being exposed to deviated content, which can be relatively complex for users to understand. This becomes more vital for users who are non-Arabic speakers and might have a wrong word translation due to its source bias. Therefore, the current study suggests that developers should use a holistic process for the authenticity of text in application development to ensure the effectiveness and adequacy of its measures. This will significantly improve the validity and reliability of the content as it is a criterion in

the initiation process of development for software. Thus, software developers can follow experts' existing literature and findings (e.g. Norman & Yasin, 2013; Almazrooie et al., 2020; Gilkar et al., 2020; Hakak et al., 2017 ,2018 2019; Islam et al., 2020).

2.5 Digital Quran Publications

Adhoni and Siddiqi (2013) report that the Quran Mobile software has been the most developed software commercial-wise. In this software, users can read Arabic text and translate. Notably, the software is installed on portable devices. Moreover, Arabic support on the device is not a requirement, which further allows non-Arab users to best better the software better early, The Quran and Hadith Portal (www.alim.org) is a social network site that emphasized Islamic content. This can include interpretations of the Quran, Hadith and historical events. Furthermore, the platform provides elements for practising the Quran for students of Isam. As a unique feature, the website provides interactive recitation of the Quran, where users can select their desired reciter, create repetition functions, view related info (e.g. interpretation or ayah), modify and change the fonts, engage in group discussions (specific to surah and ayah). As the current research introduces a new model for mobile applications using digital Quran, it is important to provide a general understanding of the status quo and various means that are available to users in this context.

Within the scope of this research, transliteration is terminology that has been used, and thus, it is defined as a corresponding character in a language for the representation of letters or words from a different language. In this sense, those non-Arabic users that seek

to use the digital Quran have different sources such as the Quran Transliteration site. The sites provide features such as reading translations of the Quran completely in several languages. Translations, basic recitation, memorization, and reading is also made available in AlMudarris Quran Software. This software enables users to have access to a wide variety of languages and search functions. In addition, it has bookmarking and note features which can be useful for non-Arab speakers. Furthermore, their platform allows verses to be copied for reference or recitation which is another desired ability for users across different devices (http://transliteration.org/Quran; Dar-us-Salam Publications (2017).

In addition to what was mentioned above, there are other platforms that provide a variety of services to users. In this sense, The Koran Mobile Application uses MP3 sounds, HTML pages and other features. Similarly, The Holy Quran Search & Live Quran tutoring at Quran Interactive (2017((http://www.Quraninteractive.com) provides a direct tutoring service covering different aspects such as Quran reading lessons, Quran reading with tajweed (recitation rules), Quran translation, Quran memorization, Qirat (reading) competition, and basic Islamic knowledge. Moreover, Pocket Quran (2017) (http://www.pocketQuran.com) is a commonly used platform across many devices. As it can be seen there are various digital sources for users, which shows the importance of the matter at hand. The current research aims to provide a pathway for optimizing storage usage which can be of aid for various platforms (Adhoni and Saddiqi, 2013).

A recent digital platform in the Malay language has been established *Surah.My: Terjemahan Al-Quran Bahasa Melayu* which allows users to read Quran with Malay

language translation (https://www.surah.my/). Alongside what was noted so far in terms of available digital sources of the Quran for users, there are specific applications that have been made for portable devices such as smartphones and tablets. Among these applications, The Palm Quran software provides a complete Arabic version of the Quran, while Pocket Quran enables users to have display functions such as *Othmanic* typeset with *Koufi* or *Naskh* fonts as well as both horizontal and vertical displays. Furthermore, it provides search capabilities for root word derivatives and highlighting. Pocket Islam (from Worldofislam.info) provides diacritical marks of hadith in Arabic alongside prayer table and schedule with Azan embedded. Furthermore, it tracks the location of the user to provide Qibla for prayer and the position of the sun. Quran Reader (from Worldofislam.info) software users can read translations of the Quran while being able to save or bookmark alongside browsing specific verses in their desired Surah (Adesina et al., 2010).

To provide a comprehensive report on the available sources and with regards to the first aim of this research, it is important to note other digital sources that are available for users and have been a point of interest for scholars in the field. Quran.com software (from Worldofislam.info) entails transliteration and introductory measures to surah as well as the English version of the texts (Adesina et al., 2010). Quran viewer from the same source, provide Quranic commentary, transliterator, indices, glossary, and search function for users as well as a plug-in for other translation that establishes a multi-lingual platform for its users. This enables users to compare different languages installed while having computer-generated Mushaf pages that exhibit original text in Arabic. Quotation software

provides search function by word and part of a word or group of words including roots, stems, and copy verses in full or partially regarding surah (Adhoni and Saddiqi, 2013).

iQuran III which is designed for iOS (iPhones and iPod touch) uses Uthmani font with color-coded pronunciation that provide enhanced readability. In addition, this software includes verse by verse translation and recitation. The Quran Recitation software has compressed AMR audio files that significantly reduce the required storage (Adhoni & Seddiqi, 2013). The Quran Majeed app enables an online search function, Arabic reading of the Quran as well as Urdu and English with the ability to bookmark pages. Search, navigation, recitation, commentary, customization, and translation for several different languages are available in Zekr Quran (Zekr – The Quran Project, Mohsen (2017).

Within the same scope, Al-Anvar provides search, comment section, indices, grouping, add-onns and various translations both offline and online. Notably, Al-Anvar is an open-source freeware (Al-Anvar & Najafian. 2017). The Android application of the Quran is open-source with indices and audio recitations that can be downloaded freely. Furthermore, the app supports sharing, bookmarks, translations, and interpretations (Quran Android 2.1.0 (2019) onwards to the last update on 15th April 2021 – versions can vary among devices at Google Play).

As the notion of current research in its first phase of conduct was to provide a thorough and comprehensive review of the existing literature, this section has shown that there are numerous services available for users in the digital Quran various formats. It is important to highlight the fact that the number is more than can fit the scope of current research (e.g. Verse by Verse Quran, and Complete Quran Site Code). However, by

introducing this software and platforms, the current research justifies its conduct as the literature clearly shows that many aspects can be enhanced, assessed, evaluated, analyzed, and implemented.

Several research works cite applications on Qur'an. These include text, like automatic text categorizations, semantic search in the Qur'an, recognition, and correction of recitations, etc. These tools and techniques show the major disciplines of Quran studies which are (e.g., Adhoni et al., 2013; Adhoni and Saddiqi, 2013):

- Reading with Tajweed (rules of recitation of the Qur'an).
- Tafseer (explanation of the Qur'anic verses).
- Translation and Transliteration of the Qur'anic verses (called ayat).
- Memorization of the verses of the Qur'an.
- Searching for verses/words of Qur'an, including semantic search.
- Qur'an Recitation and Bookmarks
- Authentication of Qur'anic verses available in various online documents.
- Speech Recognition technologies for learning Qur'anic recitations.

The latest technologies in the Quran studies have been of much interest to numerous researchers. They have reviewed the technologies in their works including Adhoni et al. (2013a & 2013b). Design, construction, implementation, and deployment of an all-around online Quran portal that is cloud-based are the main goal of researchers and developers. Accessibility of the portal and its applications are taken into account with regards to

whatever device the user may be using, be it a mobile phone, a laptop, a PDA, a tablet or a PC, to access the reading and resource areas. The content format of a digital Quran includes image, video, audio, and text are discussed in the following section.

2.5.1 Quran Text-Based Format

A text-based format called text-document watermarking is categorized into either linguistic or non-linguistic (Adesina et al., 2010). Linguistic techniques work on a document's lexical, syntactic and semantic attributes while endeavouring to maintain the meanings. Whereas, in the non-linguistic approach, modifications have to be made to the text by using different text properties, to achieve message embedding. Text-watermarking methods have been centred on shifting techniques, for instance, word-shift, line-shift, character-coding, and watermarking based on natural language. Watermarking based on natural language includes techniques for semantic transformation.

Jalil et al. (2010) categorized the methods of watermarking texts into either i) based on images, ii) syntactic modification, and iii) semantic modification methods that constitute entails substituting the initial text with by the use of newer intending to embed a message which is hidden but still holding the original ideas intended.

Semantic web technologies have been suggested as a framework for representing the Holy Quran using text preprocessing and ontology engines as shown in Figure 2.5 (Al-Khalifa et al., 2009). Subject Matter Expert (SME) mode is executed in the form of an identification tool in the manual form. The result is that the population of ontology has properties and terms. The tool process pipeline consists mainly of two parts: Arabic Text Preprocessing and Ontology Engine.

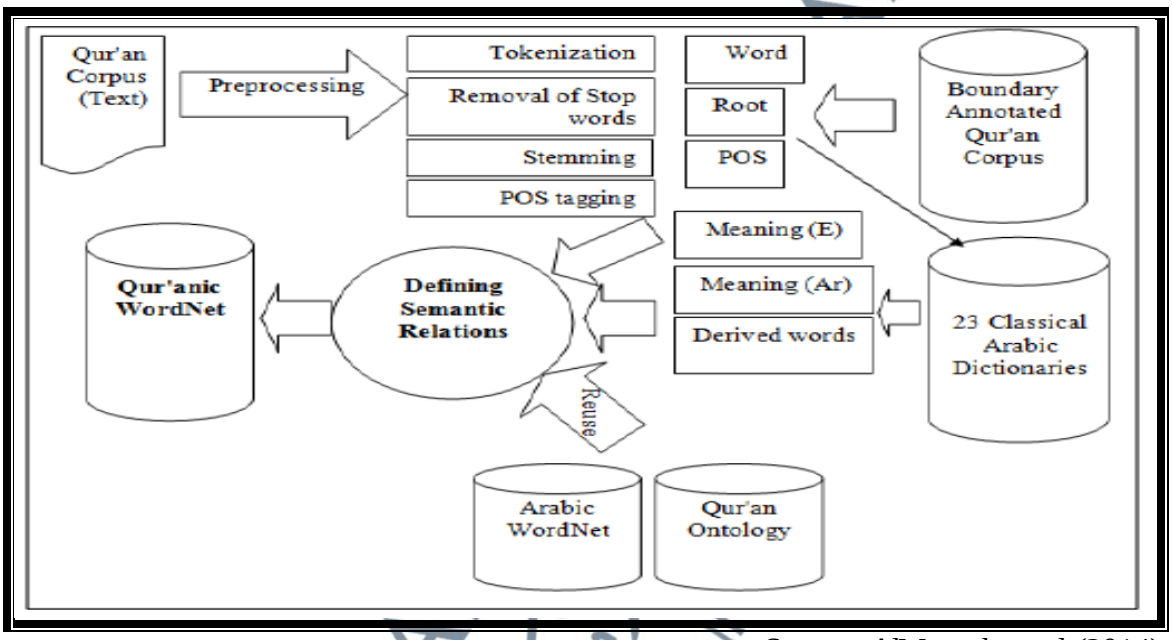
Source: Al-Khalifa et al (2009)

Figure 2.5: SemQ Tool Pipeline

The initial part known as the Verse (sentence) pre-processing includes: morphological analysis, stop words removal and Part of Speech Tagging (PoS). Part of the Speech Tagging process labels every word in a sentence with its correct tag for instance: a noun, a verb, or an adjective. For this step, Buckwalter morphology/POx Annotation is a promising tool. The stop-word removal process is essential for filtering the verse from pronouns, adverbs, and conjunctions that are not added to the semantic opposition.

The morphological analysis process is applied to locate the morphemes that are part of a word, for example, affixes and stems, with the intention that stems are the ones only outputted for the subsequent process. The next stage is the search and retrieval of its components. This is done by entering the list of stems, obtained from the previous stage of pre-processing, into the ontology engine. The engine then decides whether the semantic opposition is and establishes its degree as for whether absolute or scalar.

A new method was suggested for use in Quranic Arabic WordNet with the capability of pre-processing through stop word removal, tokenizing, POS tags, and stemming. Consequently, through the grouping of words having similar meanings, a synonym can be set (Al Maayah et al., 2014) as shown in Figure 2.6.



Source: AlMaayah et al (2014)

Figure 2.6: Quranic Arabic WordNet

An algorithm was put forth by Kamil & Jalil (2012). The algorithm could give a comparison of words in Arabic that are coded in the internal library. This was done by choosing the shortest code word possible and then encoding it with a Unicode representation, hence saving space. The Unicode is known as Romanization for the reason that an Arabic character is embedded into a Unicode in the form of 8 characters and not one character as presented in Table 2.2. this representation has been commonly used by

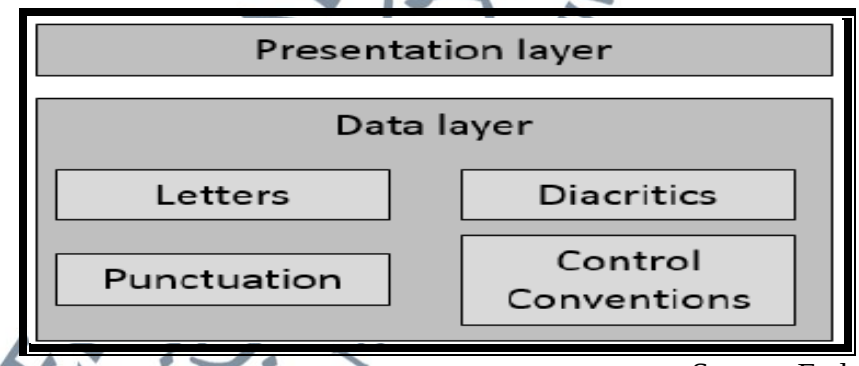
several scholars in the field (e.g. Adhoni & Seddiqi, 2013; Almazrooie et al., 2020). The current research uses the references in terms of conduct and framework.

Table 2.2: Arabic Character Representation Unicode

Character	Representation in Unicode
ي	ى
ىو	وى

Source: Kamil & Jalil (2012)

Representation of the Quran using the Quranic code was suggested by Foda et al. (2013). The code worked on character, word and phrase levels. A symbol in the Quran that did not have a Unicode was added as a new character as illustrated in Figure 2.7. The model in focus was encoded by the name of the chapter, numbered page, and the ayat chapter.

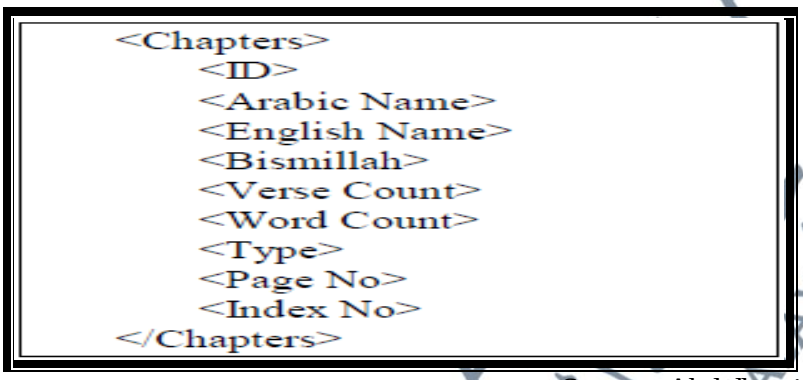


Source: Foda et al (2013)

Figure 2.7: Quranic Model

The text representation of the Quran used the same principle and was conducted by Abdelhamid et al. (2013). Figure 2.8 presents the proposed hierarchical database having

an index, pages, ID, and chapters as by researchers. The conventional method having the capability of pre-processing through stop word removal, tokenizing, POS tags, and stemming, is an application by the researchers. However, using this method on a large amount of text like the Holy Quran is very costly.



Source: Abdelhamid et al (2013)

Figure 2.8: Definition of Holy Quranic Chapters

characterized by watermarking of text documents as linguistic or non-linguistic. Linguistic strategies affect a document's lexical, syntactic, and semantic qualities while attempting to retain its meanings; however, non-linguistic ways alter the text by embedding a message utilizing various text attributes. Secondly, text watermarking strategies have included shifting techniques such as line-shift coding, word-shift coding, and feature/character coding, as well as natural-language-based techniques such as synonym replacements or language-dependent semantic transformations. Research published offered a fragile watermarking approach for preserving digital validity (Kurniawan, Khalil, Khan & Alginahi, 2014). Quran's technique is referred to as a fragile watermarking technique since it operates on the wavelet and spatial domains of digital Quran pictures. Each block of wavelet processed picture contains authentication bits. Then, the pixels' least significant bits are examined for embedding additional

authentication bits. The testing results indicate that the watermarked picture is undetectable and susceptible to common assaults.

Additionally, works on the subject of Digital Quranic Information Retrieval have been conducted using a variety of formats and methodologies, however, the majority of researchers employ standard preprocessing approaches for Quran words and verses such as stemming, tokenizing, POS tagging, and image processing. However, all of these solutions require time and storage and do not take into account the concept of duplication.

This study provides a novel approach for handling word duplication that utilizes the UTF-8 character encoding, which is backwards compatible with ASCII code and is implemented using a sparse matrix with double offset indexing. The Unicode transformation format (UTF) is the worldwide character coding standard for representing characters, whereas UTF-8 is an alternate coded representation format for all Unicode characters that maintains ASCII compatibility (Kurniawan et al., 2014).

2.5.2 Quran Image-Based Format

The work that is classified under the image-based format category is reviewed below. The discussion comprises the entirety of the work done concerning protection and verification of the integrity of the Quran in addition to the methodology used and the shortcomings met. Lots of Quran and hadith images can be found on the Internet (Hakak et al. (2017).

The image content is subdivided into two subtypes which are plain and complex images. The plain image is simply a clear picture having as few colour details as possible. On the other hand, complex images constitute pictures having additional details and many symbols incorporated into it. Figure 2.9 shows the two types of images.

Source: Hakak et al (2017)

Figure 2.9: Quranic Images Plain and Complex

There are many forms of methods that can be applied when scrutinizing the integrity of the images as this area is part of the image processing domain. Various formats such as JPEG, TIF, GIF, etc. are used to render both plain and complex images. Performance verification depends on the different techniques applied to the processing of images.

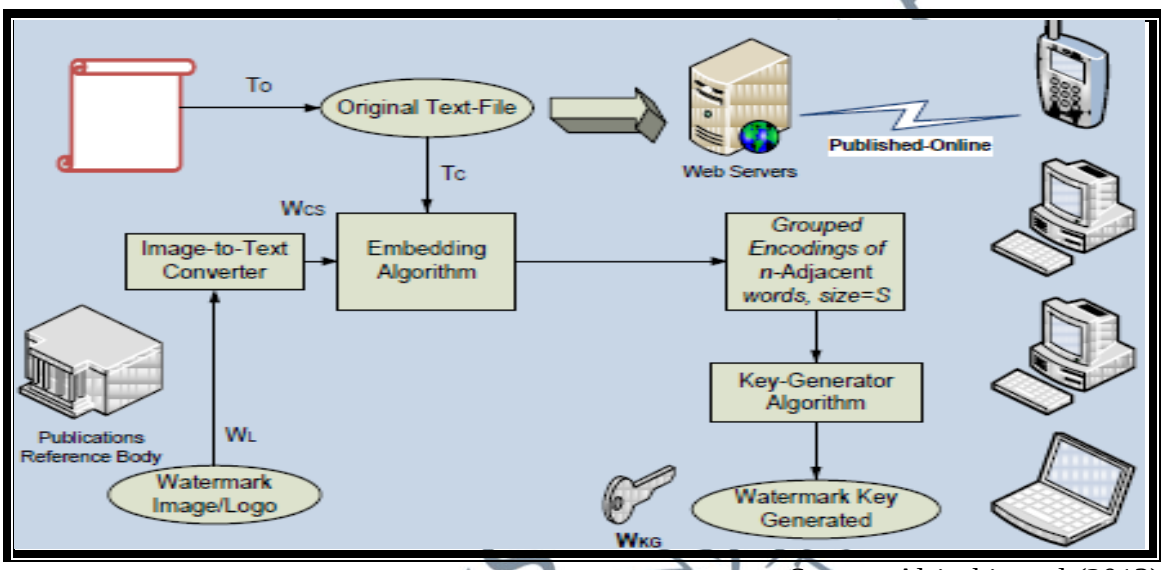
When working with online sensitive content, protection of content and copyright constitute the two significant encumbrances (Tayan et al., 2013). To verify and authenticate content, zero watermarking approaches have been seen to show potential. A

particular series of data gotten from the watermark logo is embedded into the document that is to be authenticated. Lastly, logical XOR operation is performed to create a particular key having the limit of word size.

A new method named Enhanced Singular Value Decomposition (SVD) was devised to be used to protect and authenticate the text images content that has sensitive constraints (Laouamer & Tayan, 2013). Consequently, a new technique was proposed to mitigate the challenge of protecting digital publications in transit. Various techniques have been seen to work in the transformation domain, blending in the SVD method, with the help of various methods available working in similar spaces, for instance, the Discrete-Cosine transform (DCT), the Fast-Fourier transform (FFT), and the Discrete-Wavelet transform (DWT), etc. The SVD-technique performance analysis compared to other techniques is very promising. They had shown that the watermark could be extracted almost flawlessly in many cases to several types of common attacks. Thus, it can be used easily to protect and authenticate other sensitive digital text-image content.

The available watermarking techniques have diverse abilities depending on application requirements. For instance, some techniques can detect and also localize forgery. Based on Kurniawan et al. (2014), watermarking techniques are classified into block-based and pixel-based depending on the way the watermark code is embedded into the host image. An algorithm was proposed to be used in the subdivision of the pages into text line images. Consequently, to ensure no tampering of the original content, binarization was done as a pre-processing mechanism (Nazeeh & Bany, 2015).

A new adaptive method has been put forward by Alginahi; Tayan et al. (2013) built on zero-watermarking for highly sensitive documents where verification of the content originality and authentication was done without physically changing the cover text at any rate. Figure 2.10 illustrated the process of watermark encoding.



Source: Alginahi et al (2013)

Figure 2.10: Watermark Encoding Process

2.5.3 Quran Audio / Video Based Format

There are more than 1000 audio recitations of the Holy Quran found for free online. The recitations of the Holy Quran are different from the normal reading of Arabic writings because of the special art known as "Fan al tajweed". "Fan al tajweed" is an art for the reason that not all records will recite the same verses similarly. Moreover, a director can recite the same verses in different ways because of the flexibility of the laws of Tajweed (Habash, 1986). In a study conducted by Nazeeh (2015) an algorithm was introduced to segment pages of the Quran into text line images. This process is done without any variations through the usage of the prepress method (binarization). This

representation of the Quranic code is cited by other scholars (e.g. Foda et al., 2013). These foundational studies have included character level, word level, and phrase-level with regard to characters and symbols, in particular addressing those that did not have Unicode. The current research follows the recent literature of the subject at hand to address its aims and objectives. In this sense, it is important to establish a thorough understanding of a different aspects of audio and/or video-based digital Quran.

In accordance with what was mentioned above, a considerable amount of audio files is classified and structured logically to make up the audio library that has been created. MP3 encodings which are platform-independent have been used to encode each file. Mohamed et al. (2014) presented the classification of these recordings as follows:

a. Audio recordings of the Quran's ten recitations and two narrations in each recitation.
b. Audio recordings of the five most famous and prestigious interpretations of the Quran.
c. Audio recordings of the Matan related to the various types of recitation and Tajweed to facilitate the learning and memorization of the Quran.

An Automatic Speech Recognizer (ASR) for Arabic has been developed. It was then extended to recognize the experts in the recitation of the Holy Quran (Tabbal et al., 2006).

Tabbal et al. (2006) state that a delimiter was then developed which uses the speech recognition method to extract the audio file for the Quran verses. This system is

automatic and developed using The Sphinx IV framework. The two main phases for the evaluation of the reciters' practising the Holy Quran include:

a. Preparatory phase whereby segments of speech are made to enhance the system and signal settings. The product obtained during the culmination of this preparatory phase is used as the starting point in the subsequent phase.

b. Recognizer of the Sphinx core phase uses the Hidden Markov Model as the tool for reorganization. The output from this phase is then modified into a word in Arabic. The conversion is made possible by the HashMap and breadth-first search. Beam search was included to enable the search from the dictionary database. After obtaining the identification of the combination, it will be matched to the audio verses from the file.

A new and effective technique of learning was proposed which works by the use of multimedia through the Al-Forqan technique, to facilitate learning and memorizing the effectively Holy Quran by the students (Hammza et al., 2013).

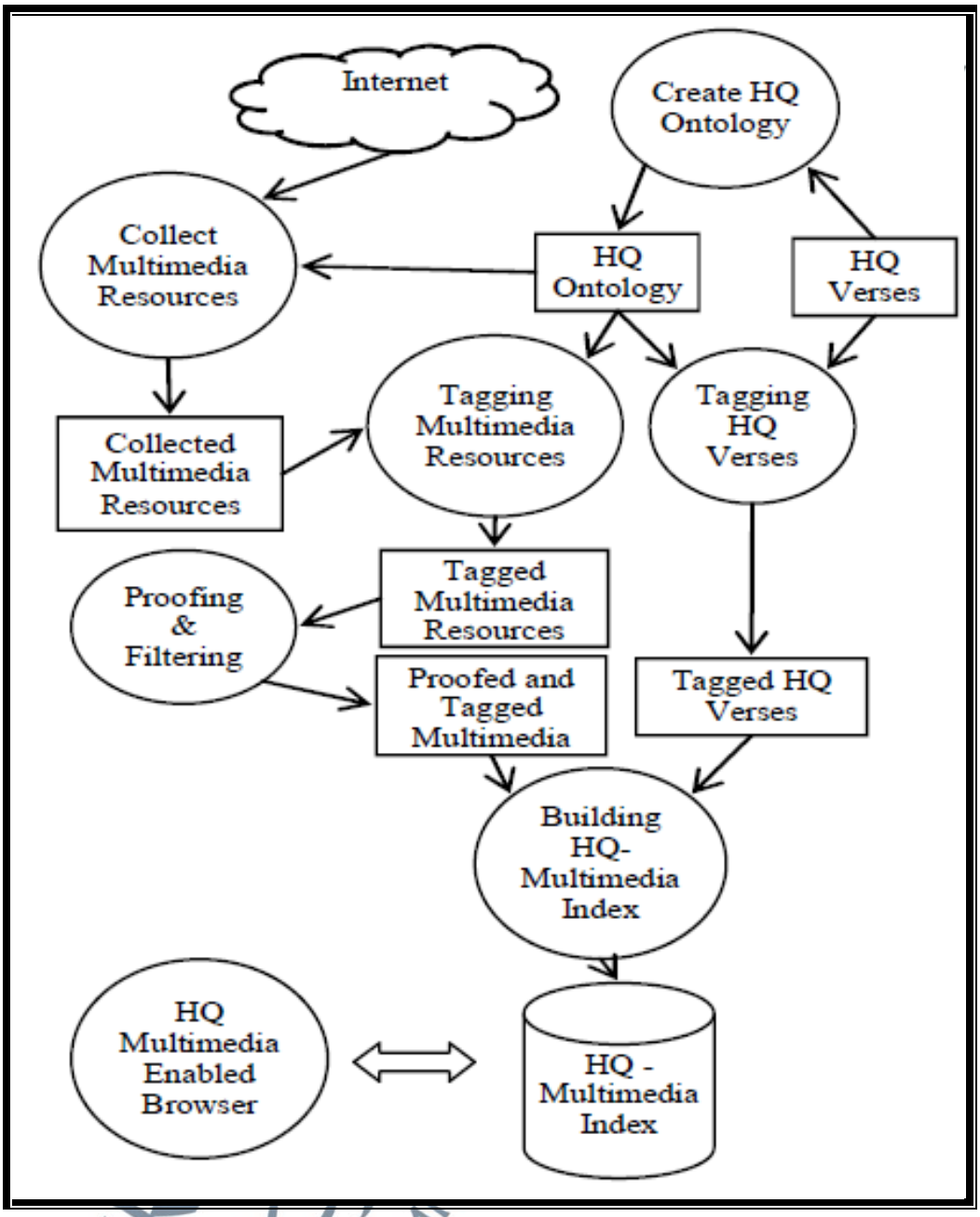
To enhance the learners' skills, motivations, attitudes, and knowledge at the same time leading to becoming skilled at reciting the Quran, a novel educational model has been introduced by Mssraty et al. (2012). This model works to facilitate the teacher in primary schools in effective teaching and recitation of the Quran. The underlying factor of this technique is the impact of interactive learning based on the use of multimedia.

A technique to determine the originality of the Quran verses was proposed by Alsmadi & Zarour, (2017). According to the authors, the two most widely used methods

are documented control and digital signature. Permission to an online document, pre and post publishing, is made possible by the use of Document control. The digital signature ensures the documents are varied by the signatory. On the other hand, ensuring accuracy for Arabic diacritics reading is a challenge as the focus is placed on integrity checking.

Hashing is also used in the research, such that, the calculation of the hash value of a particular verse is done. Thereafter, the value obtained from the calculation is compared with the one present in the database. The drawback to this method is inefficiency as the check is done one verse at a particular time. Different verses are tested using different hashing approaches.

Abdelhamid et al. (2013) suggested a system that makes available the ability of web resources to search dynamically of Quran verses with ontological terms as the underlying factors as in Figure 2.11.



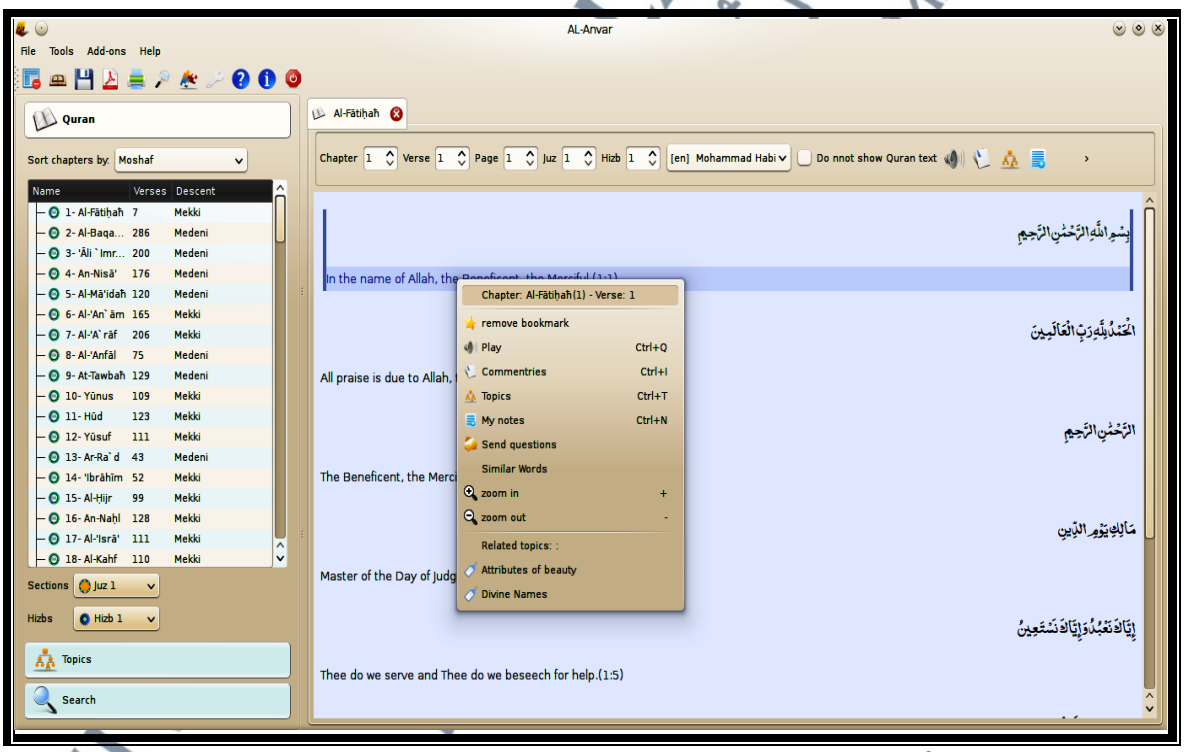
Source: Abdelhamid et al (2013)

Figure 2.11: Associating Quranic Verses with Web Multimedia Resources

When the user selects one of the resources, the software is then able to play it in audio or video form. Via the internet, a large amount of Quranic audio recording content can be accessed in the form of mp3, MPEG and mp4 files.

In the work of Subramanyam & Emmanuel (2013) different algorithms were put forward for evaluating and identifying spatial modification and temporal attacks. Bitrate, size, and the type of frame were used as Compression parameters for the detection of forgery.

Alshareef and Saddik (2012) presented a video forgery technique centered on the discovery of frame insertion and deletion illustrated in Figure 2.12.

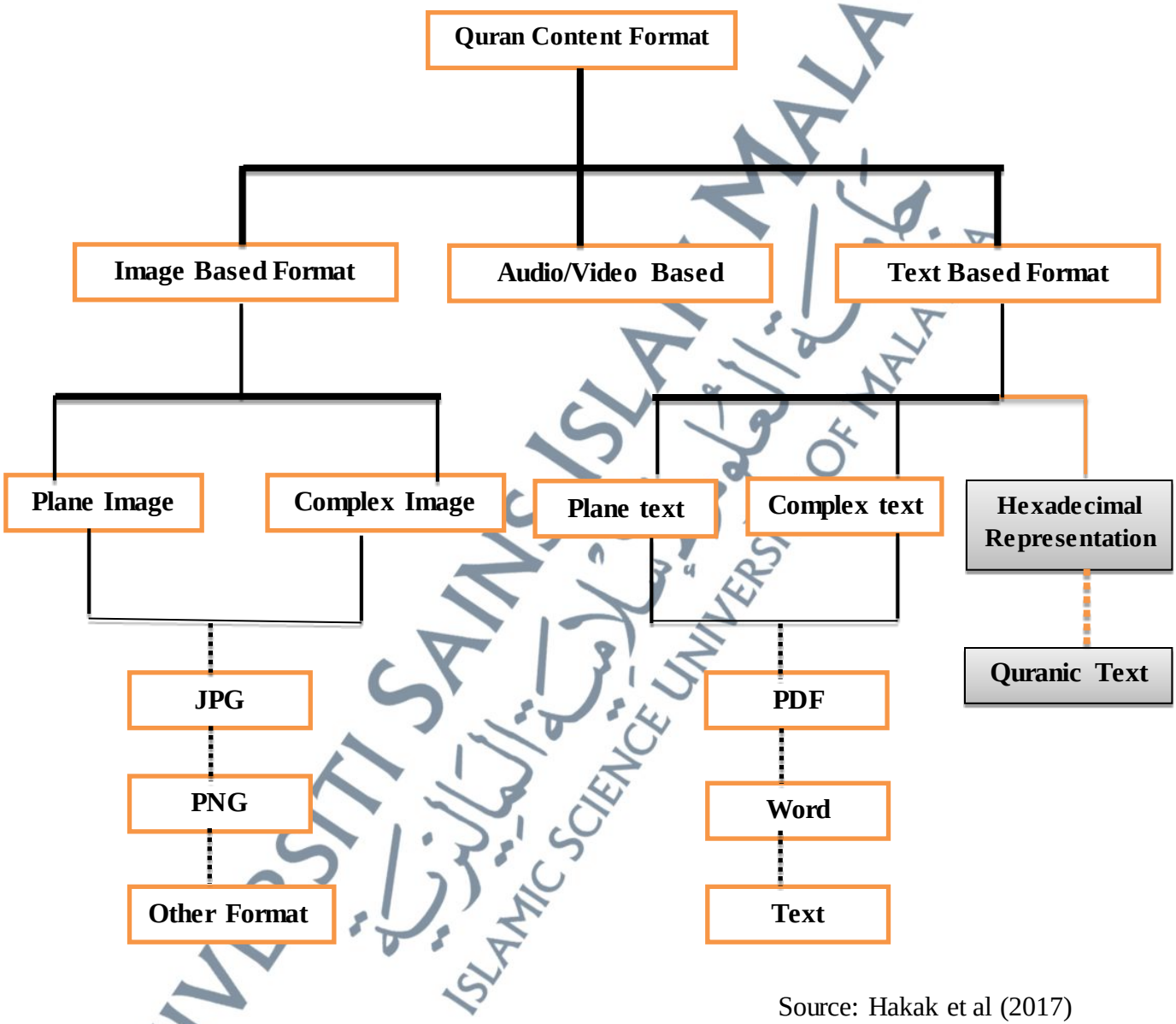


Source: Alshareef & Saddik (2015)

Figure 2.12: Al-Anvar: Quran Research Software

Hakak et al. (2017) concluded for easy access to Quran applications on mobile applications is one of the most challenging issues. Muslims around the world are downloading and following applications at a growing rate. However, there is no accurate mechanism that can verify the reliability of these applications. Hence, this issue creates a major challenge and requests for more severe analysis and research. Thus, the current research further builds on this gap to justify its conduct. This is embedded within the premise of current research as the other gaps mentioned previously in this chapter.

According to Hakak et al. (2017), the type of format used is fundamental and the classification that has been deployed in their study based on format is presented in Figure 2.13 below, which consists of an image, audio/video and text-based format. Notably, this format uses a hierarchical flow of the content to establish adequate links and funnels for data transformation and/or integration. Their work has been cited in other studies in the recent and relevant literature (e.g. Almazrooie et al., 2020; Farizuan et al., 2021; Hakak et al., 2018). Thus, another technique, using Hexadecimal Representation is proposed in this study for the representation of Quranic verses, which is a text-based approach to reduce the required memory storage as shown in the shaded box. The figure below shows the proposed model of their study, which is a theoretical framework for current research. The shaded box further illuminates the scope, in which the current research is undertaken. Accordingly, the theoretical contributions of current research to the literature can be seen in the figure below.



Source: Hakak et al (2017)

Figure 2.13: Classification Based on Digital Quran Format

2.6 Vulnerability Issues for Digital Quran

As a common means, people tend to read the Quran using a traditional printed version called Mushaf. Software developers have exerted significant efforts to satisfy the online user, and thus the gap between human and online Quran interactions has been reduced.

Quran applications on the Internet have issues regarding the reliability, functionality, and content validity of the Quran. Bundles of excellent Islamic websites have appeared across the Internet and so did many websites spread false Islamic texts of the Quran. Therefore, without stringent control and monitoring by the authority to provide standards and guidelines for Muslims using the Internet lead to many problems including the issue of unauthentic and fake copies Quran, (Shameera et al., 2017). This highlights an important matter which is the notion of authenticity in using applications, which has always been a matter of interest for scholars and practitioners in various fields due to its importance and complexity by nature. Among primary concepts of information, authenticity is data integration, which is carried out by a cryptographic hash function, implying a satisfactory level of integrity for the data that is transmitted (Almazrooie et al., 2020).

In their study, Almazrooie et al. (2020) used two different methods to address integrity verification. Cryptographic hash functions and single compression techniques which Unicode UTF-8 for Arabic characters set are used. Their findings report a significant decrease which is of vital importance in terms of theoretical and practical

implications for findings within the extant literature. Hence, the current research follows their work into establishing the parameters that are presented in the next chapter.

According to Shameera et al. (2017), there are 451 online Islamic applications, and 209 applications are digitalization of Quran applications. Converting such Quranic data into digital format is a challenging task for information systems and development-based organizations. Considering that writing is the preferred method to express ideas and share information, traditional writing has now been integrated with digital documents using certain tools, such as digital pens, digital panels, personal digital assistants (PDAs), computer hardware, and mobile phones. Most of those tools employ touch-sensitive screens, which assist the users in writing text on the screen as input to the device. However, today's online Quran and Islamic books are lagging in terms of employing structured digital content (Larsson & Hoffman, 2012). Moreover, vulnerabilities within Holy Quran mobile applications are blurry and therefore lack robustness and are prone to threats (Alsmadi & Zarzour, 2017).

Zakariah et al. (2017) proposed a pictorial representation, which classifies the Holy Quran authenticity issues into two categories. First, cryptographic algorithms and the second digital watermarking. Considering that the Holy Quran plays an important role in the daily life of Muslims, its authenticity is very important. The hard copies of the Holy Quran are printed in many Islamic countries such as Asia and Arab. Before being distributed to the local Muslims and in markets, the authenticity of the printed version is extensively checked to ensure its reliability. However, in the digital world, the use of the Internet and mobile phones have proliferated the digital version of the Quran. Numerous

versions of digital Quran applications are available on the Internet that can be freely downloaded. Since it is available for free on the Internet, the question of its reliability is raised.

Many users are concerned with the authenticity of those software applications. Since the online contents are in software form, alteration is possible using available software tampering techniques to alter the contents of the online Quran. The availability of those techniques makes the users feel inauthentic about the content published online.

2.7 Digital Quran Content Integrity

A mechanism should be developed to validate and verify the authenticity of Quranic verses, and necessary measures should be taken to avoid or detect any tampering (Alginahi et al., 2017).

Alsmadi and Zarzour (2017) presented online integrity and authentication checking for Quran's electronic version. The proposed methods adopt the hashing algorithm relying on generated decimal or hexadecimal numbers to represent words and verses and to preserve integrity and authenticity. A similar study has been presented by Kamsin et al. (2014) and Alginahi et al.,(2013c) that used Unicode centric string matching approach. Figure 2.14 exhibits the string matching approach to match or compare each string or letter from the word and verse.

Source: Alsmadi and Zarzour (2017)

Figure 2.14: Unicode Centric String Matching Approach

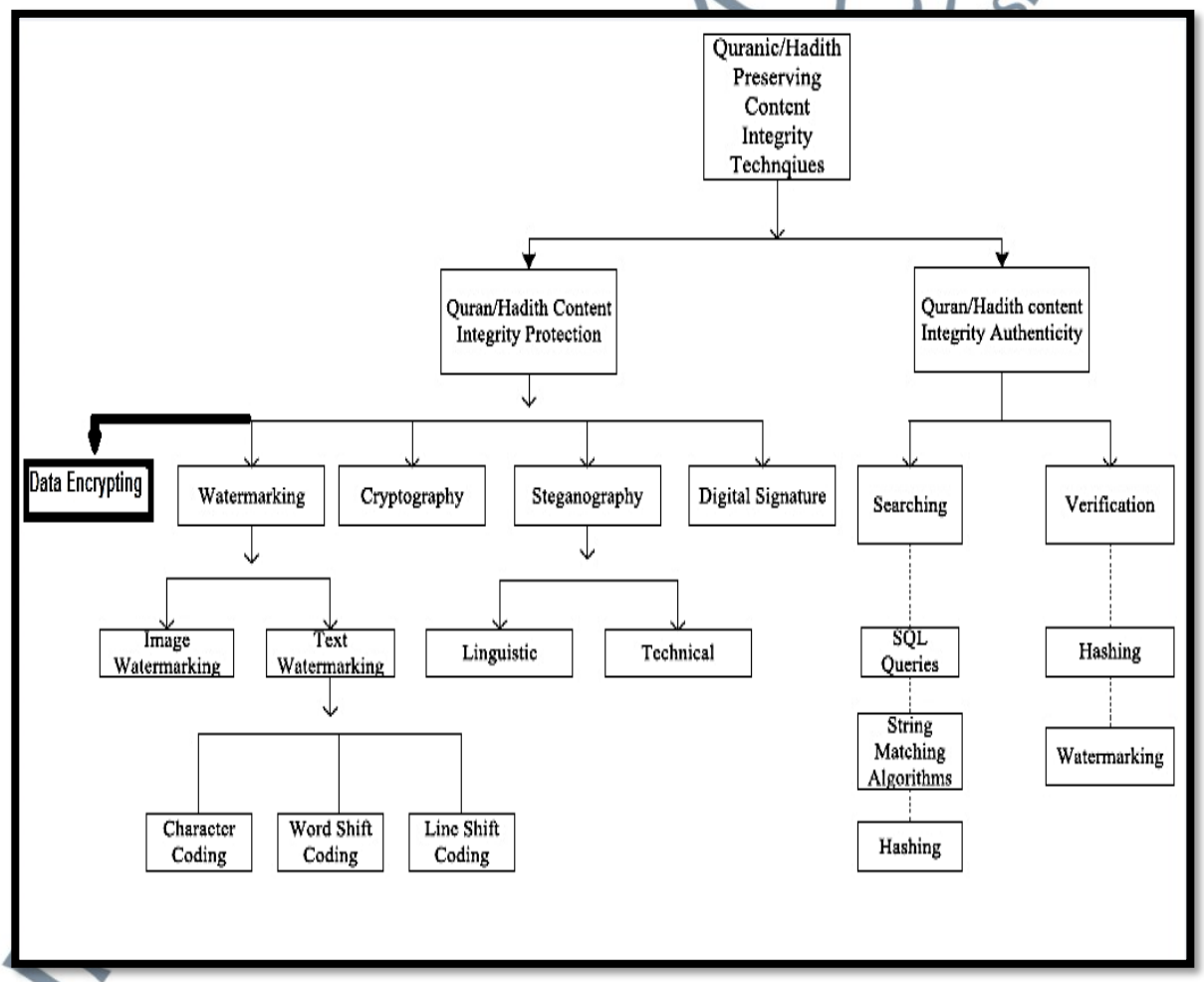
Content integrity protection is the approach where all possible techniques which are being employed for the protection of particular content or can be used for protection are put together. Hakak et. al, (2017) summarizes the protection techniques' advantages and drawbacks for image-based approaches as in Figure 2.15. Watermarking, cryptography, steganography and digital signature were the drawbacks due to high levels of attacks. This shows that this approach is more suitable for network-related attacks and less authentic for text documents and digital certification required for the sender and recipient.

Image based Approaches	Advantages	Drawbacks
Watermarking	Has the additional requirement of robustness against possible attack	Lot of attacks possible in the form of geometric, noise and other related attacks
Cryptography	Main task is to ensure users able to communicate securely over an insecure channel	More suitable for network related attacks as compared to digital documents
Steganography	Message can be sent without any suspicion. Suitable for securing transmitted messages involving encoding and decoding process	Less secure. Not suitable for preserving integrity of sensitive documents like image or text
Digital signature	very efficient in legally binding documents	Both senders and recipients need to buy digital certificates from trusted certification authorities

Source: Hakak et al (2017)

Figure 2.15: Advantages and Drawbacks of the Protection Techniques

Based on Figures 2.14 and 2.15, it can be concluded that there are challenges and weaknesses since protection techniques such as watermarking, cryptography, steganography and digital signature are less authentic. Thus, this study proposes another technique which is data encrypting as appears in the highlighted box under integrity protection in Figure 2.16 to increase the content integrity of Quranic text. This is established based on the current findings in the literature as noted in the previous section.



Source: Hakak et al **(**2017)

Figure 2.16: Taxonomy Based on Preserving Content Integrity

2.8 Related Works

In terms of authenticity, cryptography constitutes the standard technique in many applications including credit cards and banking transactions. In cryptography, the text which is readable by the human eye is converted into unreadable text or cypher text to make it inaccessible to any third or unauthorized person. Encryption, generation of keys and decryption are the three most important phases in cryptography. Encryption is the process of converting plain text into cypher text. Decryption is the reverse of encryption; Keys are used to unlocking the encryption phase (Kumar, 2016).

Quran quotes and citations found online can be checked and verified by an algorithm detailed by Alshareef and Saddik (2012). Users can verify the authenticity of the online citations of the Quran. In-depth knowledge of the features of the Arabic language and the writing style of the Quran are crucial for the algorithm whose goal is to confirm the authenticity of the Quran quotes.

A system has been proposed using the ARM microcontroller chip embedded with the ability to both plays and read the Holy Quran (Tayan et al., 2013). Another method produces a robust watermark for the image against brute-force attacks introduced by Kurniawan et al. (2013c) and Hakak et al. (2017). Watermarking is used to combine Quran images. To embed the watermark, a fragile watermarking technique is applied taking into consideration the frequency and the spatial domains. At first, discrete wavelet transformation (DWT) is applied to convert the input image to the frequency domain. The wavelet coefficients are encrypted by an authentication code enabling correlation between

blocks with embedded watermarks. This prevents attacks and enhances robustness against attacks.

A framework to determine the originality of Quranic verses obtained from internet sources such as forums and posts was presented by Sabbah & Selamat, (2013), based entirely on the assumption that obtained text verses have various diacritics. For texts having lesser diacritics, other assumptions had also been set enhancing the determination of their originality.

Digital Quran material in the form of PDF can be watermarked with another innovative technique. The method involves hashing and saves time as the hashed images have been obtained by employing the DCT algorithm. Tampering of the material can be detected by image features, which are key elements. When placed under statistical analyses, the Selected Least Significant Bits (SLSB) algorithm suffers less distortion compared to the LSB in the colour of the material while embedding the watermark (Al Ahmad et al., 2013).

A watermarking scheme that could highly improve the authenticity of the Quran by integrating artificial intelligence systems was presented by Tuncer et al. (2013). The scheme is an LSB and XOR specifically for colour images' spatial domain. LSB or XOR can detect tampering well. This is because the watermark extraction cannot succeed without the original image while performing the XOR operation.

The characters are of importance in coming up with the watermarking key. This approach is a zero-watermarking scheme. The verification authority keeps custody of the entire key generated by each Quran verse. The chapter name and number are checked

from the start. The results show that during verification, intentional or unintentional attacks and tampering could be detected 100% of the time. This was possible because the actual key for the verse was checked and verified with the one stored by the verification authority (Alginahi et al., 2013b).

Experiments have shown that high image quality could be preserved whilst achieving good results in tampering detection and localization with less watermarking (Kurniawan et al., 2014).

Refer to Table 2.3, which summarizes 16 techniques and presents the methods applied with the purpose of each method, the principal method used, and the results realized.

Table 2.3: Methods Applied in Tabular Format in Conjunction with The Purpose Of Each Method, the Principal Method Used, and Final Results Realized

Authors	Title	Aims	Method	Conclusion	Comments
Alshareef & Saddik, 2012	A Quranic quote verification algorithm for verses authentication	Develop a better framework to authenticate Quranic quotes	Quote authentication approach	Verify the Quranic e-contents over the Internet	Algorithms that discuss the Quranic quotes
Tayan et al., 2013	Quran-on-Chip (QoC): An Embedded System Framework and Design for Electronic Quran Dissemination for Internet-Enabled Devices	Authentic Quran propagation	Quran-on-Chip (QoC) subsystem within future multimedia product	Embedding the digital Quran content onto an ARM microcontroller	Compatible for embedding in other microcontroller architectures
Kurniawan et al., 2013 & Hakak et al., 2017	Diacritical Digital Quran Authentication Model & Exploiting Digital Watermarking to Preserve the Integrity of The Digital Holy Quran Images	Authentication of Holy Quran images	Fragile watermarking a method that works on block wise in the wavelet domain and pixel-wise in the spatial domain	Detect any manipulation on the content of digital Holy Quran and thus preserves its content's integrity	Public key cryptography is utilized to encrypt the authentication bits, a hash function is used
Sabbah & Selamat, 2013	A framework for Quranic verses authenticity detection in an online forum	A framework to detect and authenticate Quranic verses	Computing numerical Identifiers of words in the detected text, then comparing these identifiers with identifiers of original Quranic manuscript	The accuracy was 62% on average, while the Precision and recall were 75 and 78%, respectively	Quranic verses extracted in a text from online source especially forums posts
Ahmad et al., 2013	A New Fragile Digital Watermarking Technique for a PDF Digital Holy Quran	Watermarking PDF digital Holy Quran	Invisible fragile watermarking technique	Protecting the integrity of a PDF digital Holy Quran	DCT algorithm for feature extracting along with a Gear hash function to provide tampering detection

Tuncer et al., 2013	Watermarking application for authentication of the Holy Quran.	Authenticate the raffle and to prevent the unauthorized distribution of printed or modified in establishing the digital samples	Watermarking techniques using steganography methods, XOR, LSB, and Border watermarking techniques are used	An authentication system is developed using watermarking	"XOR Watermarking Technique" and "LSB Watermarking Technique" has been found advantageous
Alginahi et al., 2013 b	A zero-watermarking verification approach for Quranic verses in online text documents	Authentication of the Quran verses	A zero watermarking	100% detection of any distortion made intentionally or unintentionally to Quran text	A key is generated for each verse of the Quran
Kurniawan al, 2014	DWT+ LSB-based fragile watermarking method for digital Quran images	Fragile watermarking method for digital Quran image authentication and tamper identification	Discrete wavelet transform (DWT)	The watermark is authentic against local attacks	Watermark is encrypted using secret key
Sabbah & Selamat, 2014	Support vector machine-based approach for Quranic words detection in online textual content	Detecting the Quranic words in a text which are extracted from online sources	Support vector machine	Accuracy measurements achieved by the proposed approach is higher than the prior measurements	Different features Categories, such as the diacritics and statistical features are performed
Saada and Zhang, 2015	Vertical DNA Sequences Compression An algorithm Based on Hexadecimal Representation	Describe an algorithm that compresses the DNA sequence in its equivalent in hexadecimal representation	Transformation of the hexadecimal representation is followed by a conversion of the result into a binary representation	Permits an easy search of regions of similarity of a set of DNA sequences	The similarity of this approach is that it uses hexadecimal for compression and Hexadecimal but to represent only one letter

Alsmadi & Zarour (2017)	Online integrity and authentication checking for Quran electronic versions	Authentication of Quranic verses	Document control gives permission before and after publishing a document online.	A complete verse can be checked at a time. Different verses are tested using different hashing approaches	focus is on integrity checking, whose challenge lies in the correct reading of the Arabic diacritics single
Hakak et al., 2017	Preserving Content Integrity of Digital Holy Quran: Survey and Open Challenges	Systematic, analyze and categorize existing research related to preserving and verifying the content integrity of the Quran	assesses these existing studies & call for a reliable universal database of authentic and verified Digital Quran and hadith content.	Quran applications on Mobile applications are one of the most challenging issues& there is no accurate mechanism that can verify the reliability of these applications	This issue creates a major challenge and requests more severe analysis and research. & Used Unicode-centric string matching approach
Mazlan et al., 2018	Quranic Cross-Lingual Information Retrieval Optimization Using Hexadecimal Conversion Algorithm	Quranic Cross-Lingual Information Retrieval (Q-CLIR) model	Hexadecimal Conversion Algorithm by using an encoding approach	A general model for Quranic Cross-Lingual Information Retrieval (Q-CLIR) using QuHex is presented as a solution to improve the readability of natural languages	String matching approach
Almazrooie et al., 2020	Integrity verification for digital Holy Quran verses using cryptographic hash function and compression	Address integrity verification. Cryptographic hash functions and single compression techniques	Unicode UTF-8 for Arabic characters set is used.	Their findings report a significant decrease which is of vital importance in terms of theoretical and practical implications for findings.	Current research follows their work.

Golkar et al., 2020	Content Integrity Techniques for Digital Quran	Systematically identifying and categorizing suitable techniques.	Categorizing suitable techniques that can be used to preserve the content integrity of the Digital Holy Quran	Future challenges in Quran and Hadith authentication.	Content integrity can be explicitly preserved due to the sensitivity of the Quran's content
Farizuan et al., 2021	Analysis of Joc Radio FM Digital al-Quran using finite element analysis	To improve the design of the Joc Radio FM Digital al-Quran	Enhance the aesthetical value of the radio without negating product sustainability for Digital al-Quran using Finite Element Analysis (FEA)	Prove the efficiency-improved design of Finite Element Analysis (FEA) for FM Digital al-Quran	Their work has been cited in other studies in the recent and relevant literature

2.9 Character Representation for Arabic Letters

The composition of the Quran includes 77,797 words, and 6236 verses (qurananalysis.com). The vast amount of information makes it a challenge to the classification of the entire chapters of the Quran. Consequently, poor classification of chapters will make it difficult to extraction of information. Moreover, the determination of similar words, hypernym, and the meanings of the general words pose a significant representation challenge.

Arabic characters have been represented by UTF-8 character encoding having compatibility with the ASCII code in a backward manner. Arabic letter placement in a word is important because a lack of proper placement makes the Quran complex to read and recite. UTF-8 is a variable-sized coding method to encode text. The Arabic characters set are located in the code points U + 0600 to U + 06FF in the standard UTF-8 (UTF-8, 2015). Figure 2.17 shows the current UTF-8 Unicode Representation for Arabic letters encoding Arabic words. Each character needs two bytes to be coded. In the streaming data, the Arabic characters set in Hexadecimal fall into the closed interval [0xD880, 0xDBBF] (Almazrooie et al., 2020); this is illustrated in the figure below as the current research follows the same criteria for its conduct.

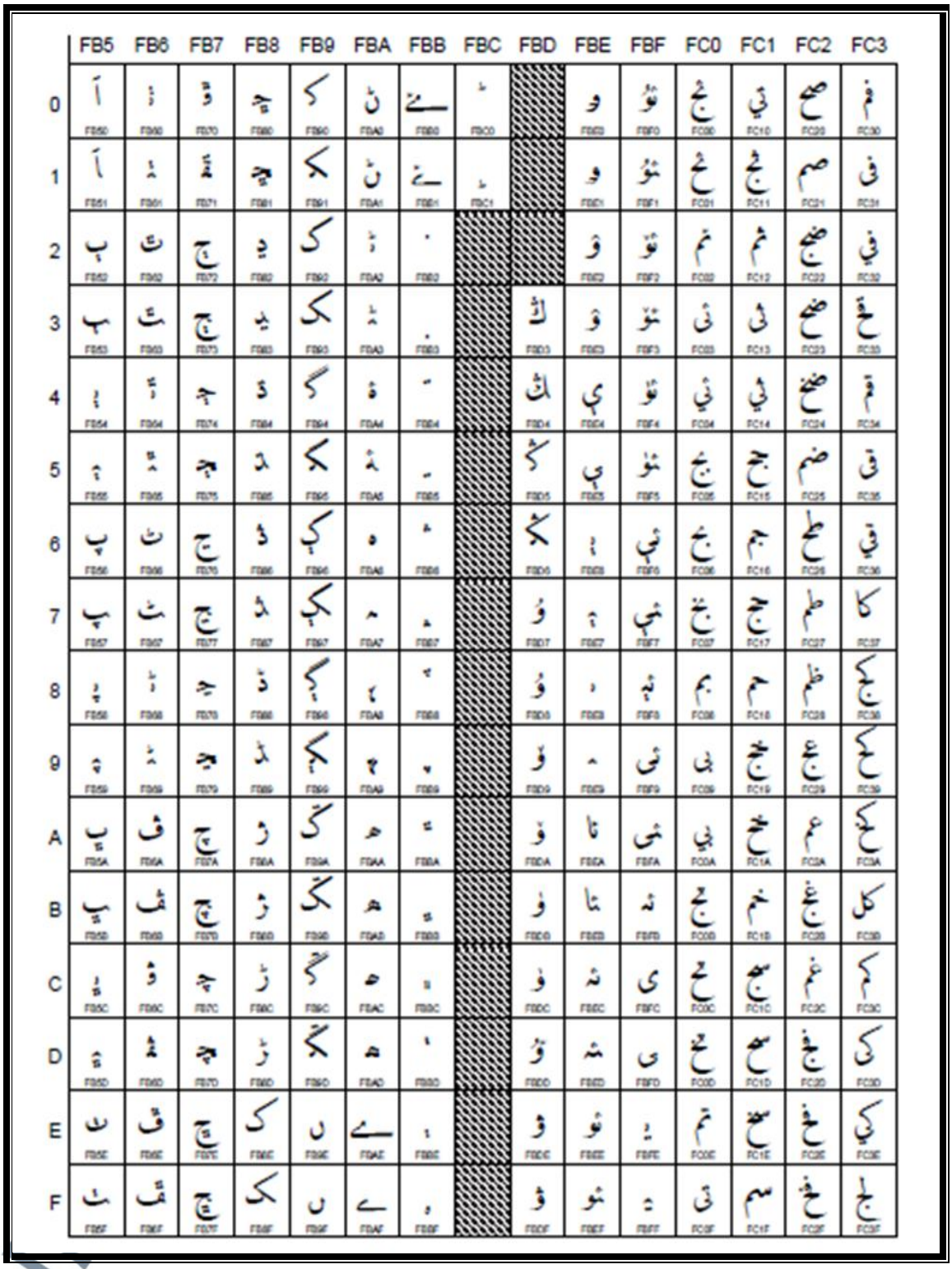


Figure 2.17: Unicode Standard 7.0, Copyright © 1991-2014 Unicode, Inc., Arabic Presentation Forms A

The following example illustrates the Thad (ض) letter in four positions having been drawn on the word. Moreover, the hexadecimal representation obtained from the Unicode Standard 7.0, Copyright © 2014 Arabic Presentation is used. According to figure 2.18 and figure 2.19.

isolated	ض	FEBD
initial	ضـ	FEBF
medial	ـضـ	FEC0
final	ـض	FEBE

Figure 2.18: Thad ض Letter in Four Expected Positions on The Word

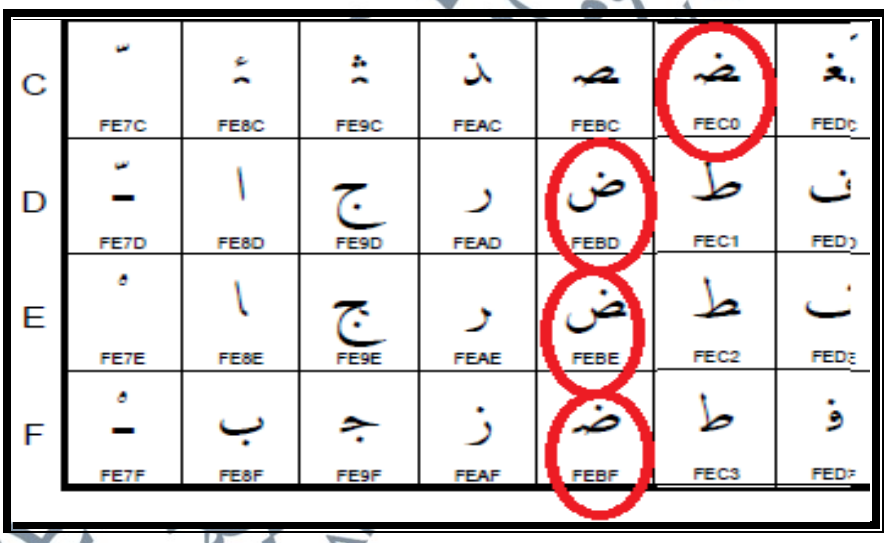


Figure 2.19: Unicode Standard 7.0, Copyright © 2014 Arabic Presentation for Thad ض Letter

The Unicode project indicates the effort towards the architectural improvement in handling text in multilingual, a technique for character encoding in computers which

allows efficient processing with the capability to cover all the languages in the world (Gupta et al., 2010).

2.10 Representation of Words and Verses in the Quran

The Quran is the primary scripture of the faith of Islam. It is the most important reference for all matters of faith, social practice, the contemplation of law and the understanding of the Divine. It is widely d as the finest work in classical Arabic literature (Jones, 1994).

The Quran is divided into 114 chapters (surah) and 6236 verses (ayah). It has been analyzed, interpreted, annotated and studied for over a thousand years. The development of computer technology made it possible to research in more advanced and powerful ways. Quranic Arabic Corpus was built as an annotated linguistic resource that shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The corpus provides three levels of analysis: morphological annotation, a syntactic treebank and a semantic ontology (corpus. Quran, 2018).

The Quranic Arabic Corpus is a collaboratively constructed linguistic resource initiated at the University of Leeds, with multi-layers of annotation, including part-of-speech tagging, morphological segmentation and syntactic analysis using dependency grammar (Liu et al., 2018).

The Quran is not only comprised of huge words but words in a repeating format. A method was proposed in the representation of Quranic words, which uses Unicode in the calculation in hexadecimal format for the individual words. The calculated word in

hexadecimal is then stored in an array. Conversion to hexadecimal is based on converting individual letters with a unique ID for each word and verse in the Holy Quran.

2.11 Data Compression Using Hexadecimal

Saada and Zhang (2015) describe an algorithm that compresses the DNA sequence in its equivalent in hexadecimal representation; it permits an easy search of regions of similarity of a set of DNA sequences. A simple subtraction operation follows the transformation of the sequences to the hexadecimal representation and conversion of the result into binary representation and detection of adjacent zero suites that represent the regions of similarity between the sequences; the algorithm is based on the binary representation of nucleotides. The similarity of this approach is that it uses hexadecimal for compression whereas Almazrooie (2020) used string and cryptographic.

Mazlan et al. (2018) proposed a qur'anic cross-lingual information retrieval optimization using hexadecimal conversion algorithm called QuHex as a potential solution to improve the readability of natural languages by using the encoding approach. QuHex utilizes the hexadecimal conversion algorithm to convert every Arabic word into its unique hexadecimal value, which was a string-matching approach to match or compare each string or letter from the word. Refer to table 2.4, which lists the characters of Arabic and Latin and code points (in hex). A similar study has been presented by Hakak et al. (2017) that used Unicode centric string-matching approach.

Table 2.4: Hexadecimal Value for Each Character.

ل	ي	خ	ن
d984	**d98a**	**d8ae**	**d986**

Source: Mazlan et al. (2018)

Almazrooie et al. (2020), regarding integrity verification for verses, assumed a collision between verses, which can be categorized as algorithm failure for the hash function. A compression method of the verses presented in Table 2.5 shows the results of compressing a Quranic diacritical verse **" الله الصمد"** using the proposed compression method.

Table 2.5: Compressing a Quranic Verse Using Proposed Compression Method

Verse	اللهُ الصَّمَدُ.
Bytes	D8A7D984D984D991D98ED987D9820FD8A7D9 84D8B5D991D98ED985D98ED8AFD98F
Compressed	274444514E474F20274435514E454E2F4F

Source: Almazrooie et al (2020)

2.12 Discussion

As noted in this chapter, various aspects within the literature have been addressed by scholars while enabling future studies such as the current research. In this study, the stance is to focus on a new representation of the Digital Quran Model that can optimize space and preserve the integrity of the Quran content on a digital platform, as shown by the traditional (sunnah) of the Rasulullah SAW companions. Since audio and video-

based representation consumes memory space, this study specifically focuses on a text-based representation, which is on the application layer as a basis of comparison towards bit and bytes-based representation or hexadecimal (presentation layer representation) (Almazrooie et al., 2020; Gilkar et al., 2020; Hakak et al., 2017, 2018, 2019; Islam et al., 2020). A recent study by Hakak et al. (2019) found that through unified approaches of watermarking and string matching methodologies content integrity can be explicitly preserved due to the sensitivity of the Quran's content.

2.13 Summary

This chapter presented a comprehensive review of literature on the following categories: digital Quran, digital Quran publications, vulnerability issues for digital Quran, content integrity, related works, character representation for Arabic letters, representation of words and verses in the Quran, data compression using hex, also highlighted the research gaps in the existing literature on Digital Quran, which could be classified into two main points; firstly, optimizing space of calculation by handling duplication of words, secondly, string representation and table representation of the digital Quran by optimizing the space according to the length of the verses and by handling duplications. The next chapter will present the research methodology of the current research.